UDC 81'322 **DOI: 10.18413/2313-8912-2024-10-2-0-4**

Kolmogorova A. V.[1]
Margolina A. V.[2]

Written vs generated text: "naturalness" as a textual and psycholinguistic category

[1] The National Research University Higher School of Economics (HSE University)
119-121 nab. Griboyedov Canal, St. Petersburg, 190068
E-mail: akolmogorova@hse.ru
ORCID: 0000-0002-6425-2050

[2] Actum CBP d.o.o. Beograd
3 Terazije, Beograd, 11000, Serbia
E-mail: avmargolina@edu.hse.ru
ORCID: 0009-0004-6180-5598

Received 05 June 2024; accepted 15 June 2024; published 30 June 2024

Acknowledgements: the article was prepared based on the materials of the project "Text as Big Data: Methods and Models of Working with Big Text Data", which is carried out within the framework of the Fundamental Research Program of the National Research University Higher School of Economics (HSE University) in 2024.

In the context of the development of text generation technologies, the opposition "naturalness − unnaturalness of text" has been transformed into a new dichotomy: "naturalness – artificiality". The aim of this article is to investigate the phenomenon of naturalness in this context from two perspectives: analyzing the linguistic characteristics of a natural text against a generated (artificial) text and systematizing introspective perceptions of Russian native speaker informants as to what a "natural" text should be like and how it should differ from a generated text. The material for the study was a parallel corpus of film reviews in Russian, consisting of two subcorpora: reviews written by people and those generated by a large language model based on prompts, which are the beginnings of reviews, from the first subcorpus. The following methods were applied for the comparative analysis of the two subcorpora: computer-assisted text processing for calculating the values of 130 metrics of text linguistic complexity, psycholinguistic experiment, expert text analysis, contrastive analysis. As a result, it was determined that from the point of view of their own linguistic characteristics, "natural" texts differ from generated texts mainly by greater flexibility of syntactic structure, allowing both omission or reduction of structures and redundancy, as well as by slightly greater lexical variability. Naturalness as a psycholinguistic category is related to the informants' autostereotypical ideas about the cognitive characteristics of people as a species. The analysis of texts erroneously attributed by informants (generated, labelled as natural and vice versa) showed that a number of characteristics of this autostereotype are overestimated by informants, while others, in general, correlate with the linguistic specificity of texts from the subcorpus of written reviews. In conclusion, we formulate definitions of naturalness as a textual and psycholinguistic category.

Keywords: Controlled generation; Naturalness; Text category; Psycholinguistic category; Metrics of text complexity; Experiment; Russian language
How to cite: Kolmogorova, A. V. and Margolina, A. V. (2024) Written vs generated text: "naturalness" as a textual and psycholinguistic category, *Research Result. Theoretical and Applied Linguistics*, 10 (2), 71-99. DOI: 10.18413/2313-8912-2024-10-2-0-4

УДК 81'322 **DOI: 10.18413/2313-8912-2024-10-2-0-4**

Колмогорова А. В.[1]
Марголина А. В.[2]

Написанный vs сгенерированный текст: «естественность» как категория текстовая и психолингвистическая

[1] Национальный исследовательский университет «Высшая школа экономики»
наб. канала Грибоедова, д. 119–121, Санкт-Петербург, 190068, Россия
E-mail: akolmogorova@hse.ru
ORCID 0000-0002-6425-2050

[2] Actum CBP d.o.o. Beograd
ул. Теразие 3, Белград, 11000, Сербия
E-mail: avmargolina@edu.hse.ru
ORCID 0009-0004-6180-5598

Статья поступила 05 июня 2024 г.; принята 15 июня 2024 г.; опубликована 30 июня 2024 г.

Информация об источниках финансирования или грантах: статья подготовлена по материалам проекта «Текст как Big Data: методы и модели работы с большими текстовыми данными», выполняемого в рамках Программы фундаментальных исследований НИУ ВШЭ в 2024 году.

Аннотация. В контексте развития технологий текстовой генерации оппозиция «естественность – неестественность текста» трансформируется в новую дихотомию: «естественность – искусственность». Цель данной статьи – исследовать феномен естественности в данном контексте с двух точек зрения: анализа лингвистических характеристик естественного текста на фоне сгенерированного (искусственного) и интроспективных представлений информантов-носителей русского языка относительно того, каким должен быть «естественный» текст, и чем он должен отличаться от сгенерированного. Материалом для исследования послужил параллельный корпус кинорецензий на русском языке, состоящий из двух подкорпусов: рецензий, написанных людьми, и сгенерированных большой языковой моделью на основе промптов, представляющих собой начала отзывов из первого подкорпуса. Для сопоставительного анализа двух подкорпусов применялись следующие методы: метод компьютерной обработки текстов для подсчета значений 130 метрик лингвистической сложности текста; метод психолингвистического эксперимента; метод экспертного анализа текста; метод сравнительно-сопоставительного анализа. В результате было определено, что с точки зрения собственных лингвистических характеристик «естественные» тексты отличаются от сгенерированных преимущественно большей гибкостью

синтаксической структуры, допускающей как пропуск или сокращение структур, так и избыточность, а также − большей лексической вариативностью. Естественность же как категория психолингвистическая связана с автостереотипными представлениями информантов о том, какими когнитивными характеристиками обладают люди как вид. Анализ ошибочно атрибутированных информантами текстов (сгенерированных, размеченных как естественные, и наоборот) показал, что ряд характеристик данного автостереотипа переоцениваются информантами, другие же, в целом, коррелируют с лингвистической спецификой текстов из подкорпуса написанных рецензий. В заключение сформулированы определения естественности как текстовой и психолингвистической категории.
Ключевые слова: Контролируемая генерация; Естественность; Текстовая категория; Психолингвистическая категория; Метрики лингвистической сложности; Эксперимент; Русский язык
Информация для цитирования: Колмогорова А. В., Марголина А. В. Написанный vs сгенерированный текст: «естественность» как категория текстовая и психолингвистическая // Научный результат. Вопросы теоретической и прикладной лингвистики. 2024. Т. 10. № 2. С. 71-99. DOI: 10.18413/2313-8912-2024-10-2-0-4

1. Introduction

Over the past decades, the theory of text elaborated a very solid repertoire of categories in terms of which any text could be described. Using the material of different languages, the researchers detail the concepts of text coherence (Wilson, 1998) and text cohesion (Dashela and Mustika, 2021; Rachmawati, Sukyadi and Samsudin, 2021), its temporality (Schramm, 1998) and informativeness.

Rapid development of AI technologies has stimulated the emergence of a new reality – artificially generated texts. The principal textual categories became not only objects for description, but also tools for generation quality assessment. Different research teams promote their own metrics to evaluate the quality of generated texts: for instance, BERTscore (Zhang et al., 2020) computes the similarity of two sentences as a sum of cosine similarities between their tokens' embeddings; Self-BLEU metric shows how diverse the output of the generated model is: 'a higher Self-BLEU score implies less diversity of the document, and more serious mode collapse of the GAN model' (Zhu et al., 2018: 4); METEOR metric (Lavie, Agarwal, 2007) and ROUGE (Lin, 2004) rely on the word overlap mechanism and compare the generated sentence to one or more human-generated reference sentences; recently elaborated metric to evaluate the structuredness of the GAPELMAPER text (Mikhaylovskiy, 2023).

However, with the growth of the performance of neural models and their widespread usage, another requirement appears: to make generated texts as natural as possible so their readers don't feel "people do not usually say it this way" or "there is something strange in this sentence". The final goal of generative models is to be able to generate texts that, when read, are perceived as being written by a human without any doubt that they were written by a human.

In such perspective, we see two amazing questions arise:

1. What is naturalness as a text category? Can we give a definition to it? If yes, it may help us, first, to "fine tune" models in a proper way to make them generate better – it means "like humans do", in a maximally anthropomorphic way;

2. What is language naturalness as an intersubjective concept that we, humans, have in our subconscious to identify ourselves as species? How do we imagine humans usually speak and what properties distinguish them, as we think, from machines, even well

trained, when they speak? This second heuristics seems to have something of magic because it refers to a very subtle substance of human attitudes, expectations and anticipations, but at the same time, this formula of naturalness, if found, will boost generation power and quality.

The present paper searches for finding answers to the research questions articulated above.

Our article is organized as follows: in Section 1, we explore the theoretical grounding for the category of naturalness; Section 2 is devoted to the overview of our methodology and data – we detail our experiments with film review controlled generation which gave us two parallel corpora of texts: actual reviews written by people and synthetical reviews generated by the Large Language Model (LLM) in response to prompts taken from actual reviews; in Section 3, we analyze both corpora comparing a) their self-BLEU score; b) their metrics of text complexity; c) results of our Human evaluation experiment focusing on the informants' subjective opinions about the naturalness of reviews – human-written and generated by AI; Section 4 summarizes our definitions of the naturalness as a linguistic and psycholinguistic category; in the Conclusion section, we finalize all our intuitions and evidence obtained in experimental work.

2. The concept of naturalness in the Humanities

Naturalness is a complex phenomenon which provokes discussions of scholars in different spheres. If we want to follow the academic tradition, we should start from the philosophical point of view.

Philosophers consider two main approaches to investigate naturalness: to see it in a moral light (in situations where naturalness is morally relevant or in some other sense provides an important criterion for decision making) or as a gradient category (in cases where something could be evaluated as more natural than something else) (Siipi, 2008). No doubt, it is the last approach that is of interest for us. Within it, researchers distinguish three types of natural forms: historically-based, property-based and relation-based. If the first type could be omitted, the last two seem very productive for our further discussion. Property-based forms appear when entities are found natural or unnatural because of their current properties or features. As we shall see, when linguists and translation professionals consider the concept of naturalness, it is precisely these forms that they mean. Relation-based forms emerge when people tend to consider those entities to which they are accustomed and which occur relatively frequently to be natural. We will adopt this point of view when analyzing naturalness as an attitude of informants when they assess the naturalness of generated texts while evaluating the quality of generation.

Naturalness as a property-based category first attracted the attention of experts in translation studies. This is not surprising in the sense that the main task of the translator is to make sure that the recipient of the text does not have a feeling that the text is foreign, that it is alien to the mode of textual production that is appropriate in his or her native community.

However, the question is not easy to answer – researchers' views differ. M. Rogers bases his argumentation on the feeling of a lack of "naturalness", which he defines as "something in the source text that is not yet adequately transparent in the target text: a particular word, a cohesive device (or lack thereof), a collocation, a distribution of information" (Rogers, 1998: 10). P. Newmark links naturalness with the concept of settings of the text – situations, contexts where the text is typically published or found (Newmark, 1987); L. Venuti − to the concept of "fluency" and "invisibility", which suggests that naturalness is characterized by being modern, widely used, standard, and consistent in its variety (Venuti, 1995). More pragmatically, naturalness is seen as the result of communicative or free translation strategies used by the translator (Serce, 2014)

or as a phenomenon being permanently in tension with the translation accuracy category (Obeidat et al., 2020).

In linguistics, perhaps, the first to speak of naturalness as a characteristic of language was Ch. Bally. Calling linguists to study not the sterile correct language of classical texts, but real speech, he defined this very real speech as an expression of our natural desire to live, of our vital energy. He wrote: “Since life is not driven by pure ideas, the language that expresses them cannot be a logical creation. In contrast to organized language, which is intellectual and logical, an affective language emerges, which is like the vital principle of language. It’s constantly at work, because in contact with life, even the most apparently objective ideas are impregnated with affectivity and become value judgments” (Bally, 1913: 25).

In other words, for Bally, naturalness means the illogicality, the emotionality that the encounter with real life brings to the strict canvas of language.

Many years later, J. Sinclar, seeking to refine naturalness in the way it could be measured, proposed three scales: neutrality − isolation – idiomaticity. A natural sentence occupies a middle range of each of these: not very neutral, but not overly affective; not completely isolated, but not very text-dependent; not sufficiently obvious to sustain any variation, but not exposed to unmotivated variation (Sinclair, 1983).

Accepting the idea that naturalness is a relative category, Orešnik (2002: 143-145) offers his own list of language naturalness basic characteristics that are available for scaling: the principle of least effort (processing ease), prototypicality, cognitive simplicity, and relative frequency of “natural” items and constructions.

In computer science, when willing to evaluate the naturalness of texts generated by the model, researchers invite assessors to participate in the so called Human Test. The participants are asked to answer a question: “Is this sentence (or utterance, or text) natural?” (Novikova et al., 2016) based on their intuition. Not rare are the experiments based on the Likert scale rating: the informants are provided with a sentence under evaluation and a five-points scale where 1 means unnatural and 5 – sounds very naturally. They are asked to rate the naturalness of the sentence. In such tasks, while doing assessment, people react introspectively based on feelings and attitudes formed in their everyday experience.

To summarise, firstly, the category of naturalness as defined by linguists remains in fact somewhat invisible - it is something that language users do not recognise as something marked. Considered by professionals in computer science in a very pragmatic way, naturalness looks like an attitude to species-specific features of humans felt by themselves.

In our paper, trying to find an answer to the question of what naturalness is, we obtained two corpora of texts − a corpus of written film reviews and a corpus of film reviews generated on the basis of prompts from the first corpus. We then sequentially applied 3 types of analyses to this material: 1) automatic metrics used to assess generation quality in computer science, 2) linguistic metrics commonly used to assess text complexity, and 3) a psycholinguistic experiment involving assessors who manually assessed text naturalness.

3. Aim of the study

The aim of this study is therefore to provide a possible answer to the question of what is naturalness as a category of text?

A priori, we assume that naturalness denotes something that is intrinsic to humans, something that derives from human nature. In other words, our goal is to characterize human-written review texts as natural per se against the background of generated texts by comparing them on three levels of analysis: text diversity measured by cosine distance between adjacent sentence embeddings, text linguistic complexity measured automatically by 130 metrics, and text subjective evaluation made in a psycholinguistic experiment. The features found in written texts and missing in

generated texts are interpreted as naturalness features, which we use to categorise naturalness.

Our hypothesis could be formulated as follows: if naturalness is actually a relevant human- written text property, it must be "palpable" either on the level of formal metric used to evaluate text generation quality; or – on the level of linguistic metrics of text complexity; or – on the level of language users' intuition.

4. Data and methodology

4.1. Data

In order to generate the data, we have finetuned the ruGPT3[1] language model on a dataset of cinema reviews for controlled text generation with a choice of sentiment. We have focused our attention on reviews for three reasons: 1) such texts are short and their small size ensures fast and high-quality training of the LMM without using very large computing power; 2) they belong to the genre of opinion texts, tending to express subjectivity, affectivity and personality – in this way, they are very revealing when we deal with the dilemma "natural vs artificial" and their inconsistency in style, flippancy in language etc. are very "useful" when searching for "human likeness" ; 3) film reviews express sentiment, emotions and, by using them as data to train model, we complicated the generation task – we wanted to test naturalness in emotionally saturated texts.

The original dataset of 200,000 reviews with three types of sentiment (positive, neutral, and negative) was collected from a Russian-language web platform by using the Selenium library in Python. The negativeness or positiveness of the review is marked by special emoji when a review is published on the platform. As negative opinion always prevails, to balance the sentiment distribution, we set a limit of 20,000 reviews for each sentiment (Margolina, 2022; Kolmogorova, Margolina, 2023).

To set up ruGPT language model we didn't apply routine steps of text preprocessing (converting text to lowercase, removing special characters and stop words, or lemmatization) because such data cleaning can significantly reduce the accuracy of the transformer model (Alzahrani and Jololian, 2021). The only stage of preprocessing was tokenization.

We have trained three models: ruGPT3Large, P-tuned (Liu et al., 2022) model and a model trained on the basis of the methodology of the Transformers Reinforcement Learning (Mnih et al., 2015). As the first of the three models mentioned had performed best, we used data from it in our further experiments.

A simple method for addressing the style transfer task involves performing a basic fine-tune of a LMM. "This approach represents a novel contribution to the field of style transfer, offering a new, yet simple means of leveraging LMMs for stylistic manipulation of text data: fine-tuning of LMM on the dataset with included prompt in it" (Margolina, 2022: 16).

In order to input the data into the model, a transformation was applied to convert the original csv table format into a plain textual file format (.txt) with a prompt:

[<s>Sentiment: [positive, neutral, negative]\nText:[text of the review]</s>]

Incorporating reviews with a prompt into the model serves to facilitate the memorization of patterns by the ruGPT3 (Li et al, 2022). This is achieved by utilizing the second segment of the prompt, which serves as a continuation that the model must generate, namely, the review itself.

The training parameters influence the output and the quality of the fine-tuned LLM (Table 1). The parameters were chosen to fit ruGPTLarge on the home GPU memory (NVIDIA GeForce RTX 4090, 24 GB memory). We chose the minimal batch size – 1 and the learning rate is the default one. With these settings and GPU, the large model took 6 and a half hours to be finetuned.

[1] Sber Devices (2021). Hugging Face: ruGPT3Large, available at https://huggingface.co/ai-forever/rugpt3large_based_on_gpt2 (Accessed 24 March 2024).

Table 1. ruGPT3Large fine-tuning parameters and ruGPT3Large generation parameters
Таблица 1. Параметры «тонкой настройки» модели ruGPT3Large и параметры генерации модели ruGPT3Large (Margolina, 2022)

ruGPT3Large fine-tuning		**ruGPT3Large generation**	
Parameter	**Value**	**Parameter**	**Value**
num_train_epochs	5	repetition_penalty	5.0
per_device_train_batch_size	1	top_p	0.95
per_device_eval_batch_size	1	Top_k	5
block_size	1024	temperature	1
lerning_rate	2.5e-4	no_repeat_ngram_size	2

As for the generation parameters (Table 1), to make our model write more freely and less trivial, we set temperature = 1.

The repetition penalty parameter was adjusted to five in order to decrease frequent repetitions. No_repeat_ngram_size restricts the appearance of repeated bigrams in the text, further promoting diversity (Table 1).

Top_k parameter removes the least probable tokens below the k-token, improving the quality of text by removing off-topic words. As in some cases, there is a huge pool of relatively possible words, top k sampling makes the text less realistic and diverse. To address this issue, we applied top_p sampling (Holtzman et al., 2019), which consists in selecting 'the highest probability tokens whose cumulative probability mass exceeds the pre-chosen threshold p' (ibid., 5). This technique permits to avoid sampling extremely unrealistic tokens, but it preserves diversity when the highest-scoring tokens have low confidence.

Finally, early stopping can be used to prevent the model from generating endless or very long sequences. This is achieved by setting the parameter to stop the beam search when a certain number of sentences are finished per batch.

Table 2. presents a sample of the ruGPT3's sentiment-controlled text generation results in Russian for three distinct sentiments.

Table 2. Example of results obtained in sentiment-controlled text generation
Таблица 2. Пример полученных результатов контролируемой генерации по параметру сентимента

Sentiment	**Generated reviews**
neutral	заголовок моей рецензии – цитата из песни группы Paramore 'Love me again', которая является саундтреком к фильму Сумерки. Фильм снят по одноимённому роману американской писательницы Стефани Майер, который был экранизирован компанией 'Warner Bros.' в 2009 году. Сюжет повествует нам о девушке Белле Свон, которая переезжает жить к своему отцу в небольшой городок Форкс. Там она знакомится с загадочным Эдвардом Калленом, у которого есть необычная способность превращаться в летучих мышей. После недолгих отношений Эдвард спасает Беллу от похитителя Джейкоба Блэка, после чего между ними вспыхивает страстная любовь. Актёрская игра Кристен Стюарт мне понравилась. Она хорошо справилась со своей ролью молодой девушки, влюблённой в вампира. Роберт Паттинсон исполнил роль Эдварда Каллена. На мой взгляд,

	ему удалось воплотить образ таинственного [The title of my review is a quote from Paramore's song 'Love me again', which is the soundtrack to the film Twilight. The film is based on the novel of the same name by American writer Stephanie Meyer, which was screened by Warner Bros. in 2009. The plot tells us about a girl Bella Swan, who moves to live with her father in the small town of Forks. There she meets the mysterious Edward Cullen, who has an unusual ability to turn into bats. After a short relationship Edward saves Bella from the kidnapper Jacob Black, after which passionate love erupts between them. I liked Kristen Stewart's acting. She coped well with her role of a young girl in love with a vampire. Robert Pattinson played the role of Edward Cullen. In my opinion, he succeeded in embodying the image of a mysterious]
negative	Итак, сперва плюсы - визуальная составляющая (которая, к слову сказать, на высоте) – тут придраться не к чему. Снято все очень красиво и качественно; спецэффекты в некоторых сценах просто поражают своей масштабностью и детализацией (чего только стоит момент с подъемными кранами). Саундтрек также порадовал необычным сочетанием классики рока и современных хитов. В общем, если вы хотите приятно провести время за просмотром хорошего фильма, то 'Аватар' вам определенно подойдет. Минусы же заключаются в том, что это далеко не самый лучший фильм из тех, которые я когда-либо видела. И дело даже не в банальном сюжете или плохом воплощении задумки режиссера. Дело в абсолютно бездарном подходе к созданию кинофильма. Начнем с сюжета. Сюжет настолько примитивен, насколько это вообще можно себе представить. На протяжении всего фильма меня не покидало чувство дежавю. Я уже где-то это видела... [So, first of all, the pluses - visual component (which, by the way, is at the height) – there is nothing to quibble with. Everything is filmed very beautifully and qualitatively; special effects in some scenes simply amaze with their scale and detailing (what only the moment with cranes is worth). The soundtrack also pleased with an unusual combination of classic rock and modern hits. In general, if you want to spend a pleasant time watching a good film, then 'Avatar' will definitely suit you. The downside is that it is far from being the best film I've ever seen. And it's not even about the banal plot or poor implementation of the director's idea. It's about the absolutely mediocre approach to making a film. Let's start with the plot. The plot is as primitive as it can be imagined. Throughout the film I had a feeling of deja vu. I've seen it somewhere before...]
positive	Этот фильм заставляет задуматься о таких важных вещах как семья, дружба, взаимопомощь и конечно же любовь к ближнему своему. Актёрская игра просто великолепна. Каждый актёр идеально подходит для своей роли. Особенно хочется отметить Джона Кьюсака сыгравшего роль главного героя Теодора Твомбли (Теодор – главный герой фильма). На мой взгляд это одна из лучших его ролей за всю его карьеру. [This film makes you think about such important things as family, friendship, mutual help and, of course, love for your neighbour. The acting

	is just great. Each actor is perfect for his role. Especially I would like to note John Cusack who played the role of the main character Theodore Twombly (Theodore is the main character of the film). In my opinion, this is one of the best roles of his career.]

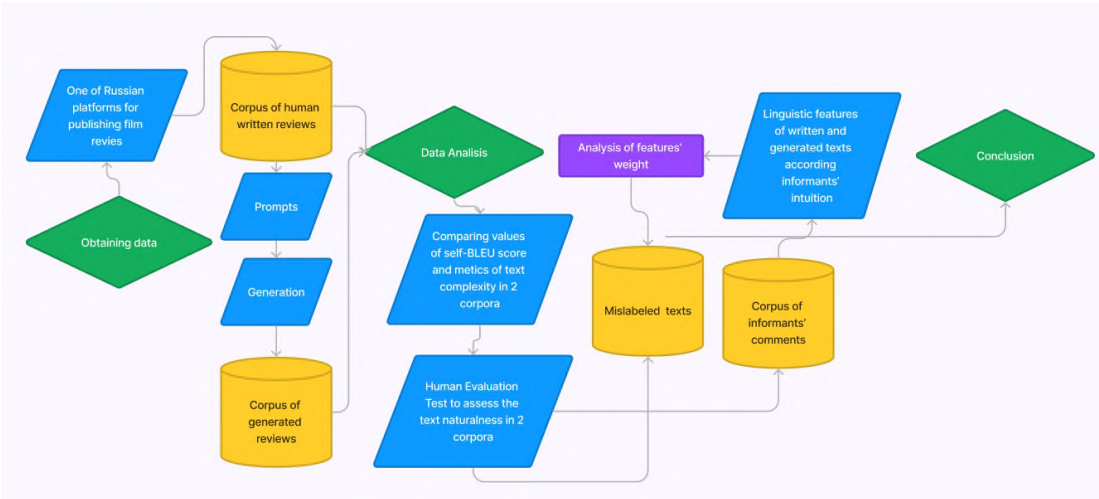
In our further experiments we use four samples: 1) a sample of 1190 human-written reviews from a well-known Russian film review platform; 2) a sample of 1190 reviews generated by the ruGPT3Large model using the prompts from the first sample and controlled by the parameter of their sentiment (like those that are shown in the Table 3); 3) a sample mixed from two mentioned above corpora and consisting of 36 reviews (18 generated and 18 written) exposed to the Human Evaluation Test; 4) a sample of 126 comments given by informants when answering about introspective intuition they were guided by while labeling texts as written or generated.

4.2. Methodology

Two of our subcorpora of 1190 texts are compared twice: firstly, by using metrics to evaluate generation quality; secondly, by exposing them to a number of metrics of linguistic text complexity. A mixed subcorpus consisting partially of generated reviews and partially of written ones was exposed to texts naturalness assessment in the Human evaluation test (130 informants participated).

As a result, we obtained a restrained subcorpus of texts wrongly attributed by assessors (synthetical texts – to written ones and vice versa) and a collection of informants' comments demasking their introspective intuition they were guided by while labeling. We grouped generated text features and written text features introspectively indicated in informants' comments into 1) syntactical, 2) stylistic, 3) lexical features, and 4) features manifesting themselves on the level of textual categories, and 5) on the level of text affectivity. Then, we did a linguistic analysis of the pool of mislabeled texts according to the set of linguistic features built by informants previously. To make the logic of our work more explicit, we visualize our workflow (Figure 1).

Figure 1. Research workflow
Рисунок 1. План исследования



4.2.1. Metrics to evaluate generation quality

To evaluate formally the quality of obtained generation, we used two well-known metrics: BERTscore and self-BLEU.

BERTscore, introduced in (Zhang et al, 2020), is an evaluation metric widely applicable in text summarization, machine translation, and text simplification. For example, in our case it calculates contextualized text embeddings and gauge the semantic similarity from 0 to 1 between model-generated reviews and the original reviews prompted by the same input. The model presumes that 1 signifies complete textual similarity (every word is repeated in both sequences) and 0 indicates no identical words between the two sequences.

Another metric, the self-BLEU score (Zhou & Bha, 2021), measures the diversity of generated text by comparing sentences within the text. It is calculated by treating each generated sentence of one sample as a "reference" and comparing it to the other using the BLEU (Bilingual Evaluation Understudy) score, which measures similarity between two texts. Lower self-BLEU scores indicate greater diversity in the generated text.

By considering self-BLEU scores, we can evaluate whether the model can produce diverse sentences while upholding quality and coherence. This evaluation aids in identifying potential overfitting issues, as a high self-BLEU score may suggest that the model is generating repetitive or clichéd outputs, potentially limiting its applicability in real communication (Celikyilmaz, 2021).

4.2.2. Text complexity assessment

To feature human-written texts and their synthetical counterparts on purely linguistic level, we use a computer model elaborated by O. Blinova and N. Tarasov (Blinova, Tarasov, 2022). It covers all existing text complexity metrics. However, we rewrote the code to run the model with our data.

The texts of two subcorpora (written and generated) were subjected to such computational analysis in parallel.

4.2.3. Human evaluation test and expert analysis

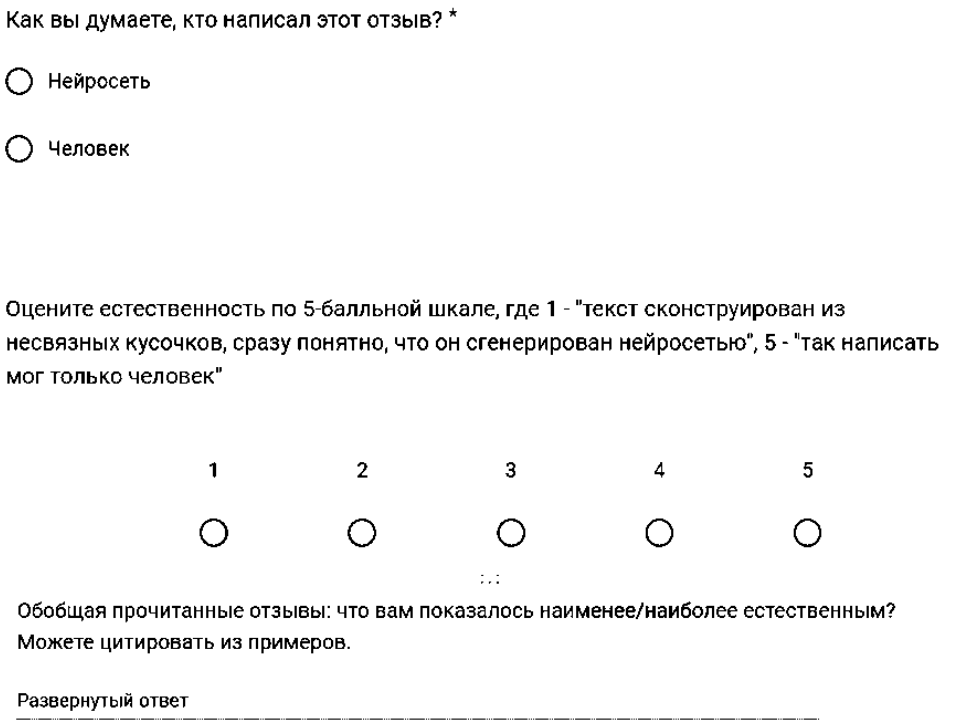
To access the subjective perception of the degree of texts naturalness, we applied another method – the method of psycholinguistic experiment.

Our pool of respondents consisted of 130 (mean age=21.3) students of the Department of Philology of HSE University in Saint-Petersburg. For us, it was crucial to invite people sensible to text quality, to text style. The respondents were recruited via an open call published in student publics in the social network VKontakte. Via Google form service each of the informants was invited to read 36 texts (one by one) and 1) to hypothesize what kind of text they read *(Who is the author of the review? – AI or Human*); 2) to assess the naturalness of the text on a five-point scale (a score of 1 denotes entirely artificial text, while a score of 5 signifies text that appears as if exclusively crafted by a human writer); 3) to formulate markers they had used to distinguish "natural" text from generated one. The interface is presented in Figure 2. The informants had no limit of time while working with 36 texts and our tasks. Each informant was working from his/her home computer.

We intentionally did not use the structured survey method. Firstly, we could not offer the informants a list of criteria by which they could evaluate the naturalness of the text as these have not yet been developed in the NLP scientific community. Secondly, there is an established tradition of the Human Evaluation Test to choose between explanatory or confirmatory questions (Schuff et al., 2023).

We focused on the first one, because "in the case of an exploratory research question, an experiment should be designed to collect initial evidence which can then be used to generate post hoc hypotheses" (Ibid: 5).

Figure 2. Interface of Human Evaluation Test on Google Forms
Рисунок 2. Интерфейс для проведения психолингвистического эксперимента



After having collected all statistics of human evaluation, we selected generated texts wrongly categorized as human-written and vice versa.

To determine whether the informants were really relying on the introspectively formulated features of a human-generated or written (in our terminology, "natural") review text, while assessing texts in experiment, we invited experts-linguists (N=3). They were given the task of identifying the characteristics of the informants in texts that had been incorrectly rated by the informants.

5. Results

5.1. Generation quality evaluation

As we can see (Table 3), the Fine-tuned ruGPT3Large model shows a rather high value of BERTscore of 0.674, which indicates that the model generates a new text closely resembling the reference text.

To test the self-BLEU score we also put in comparison our dataset of human-written reviews.

Table 3. BERTscore and self-BLEU counting results
Таблица 3. Результаты подсчета метрик BERTscore и self-BLEU

Metric	Human	ruGPT3Large Fine-Tuned
BERT-score F1	*	0.674 (max=1)
self-BLEU score mean value	0.074323	0.032231

The findings presented in Table 3 demonstrate that the Fine-tuned ruGPT3Large model attains the lowest average self-BLEU score, indicating its ability to produce highly diverse texts. Interestingly, both the human-written and generated texts display substantial diversity, with the metric values slightly favoring the neural-generated texts over the human-written ones.

Thus, formal generation quality metrics did not say anything about the naturalness of generated text. Moreover, the self-BLEU

score shows that the LLM slightly outperforms the human ability to write diverse texts. In this regard, we decided to resort to linguistic analysis and psycholinguistic experiment. The results of both procedures will be described in the following parts of the article.

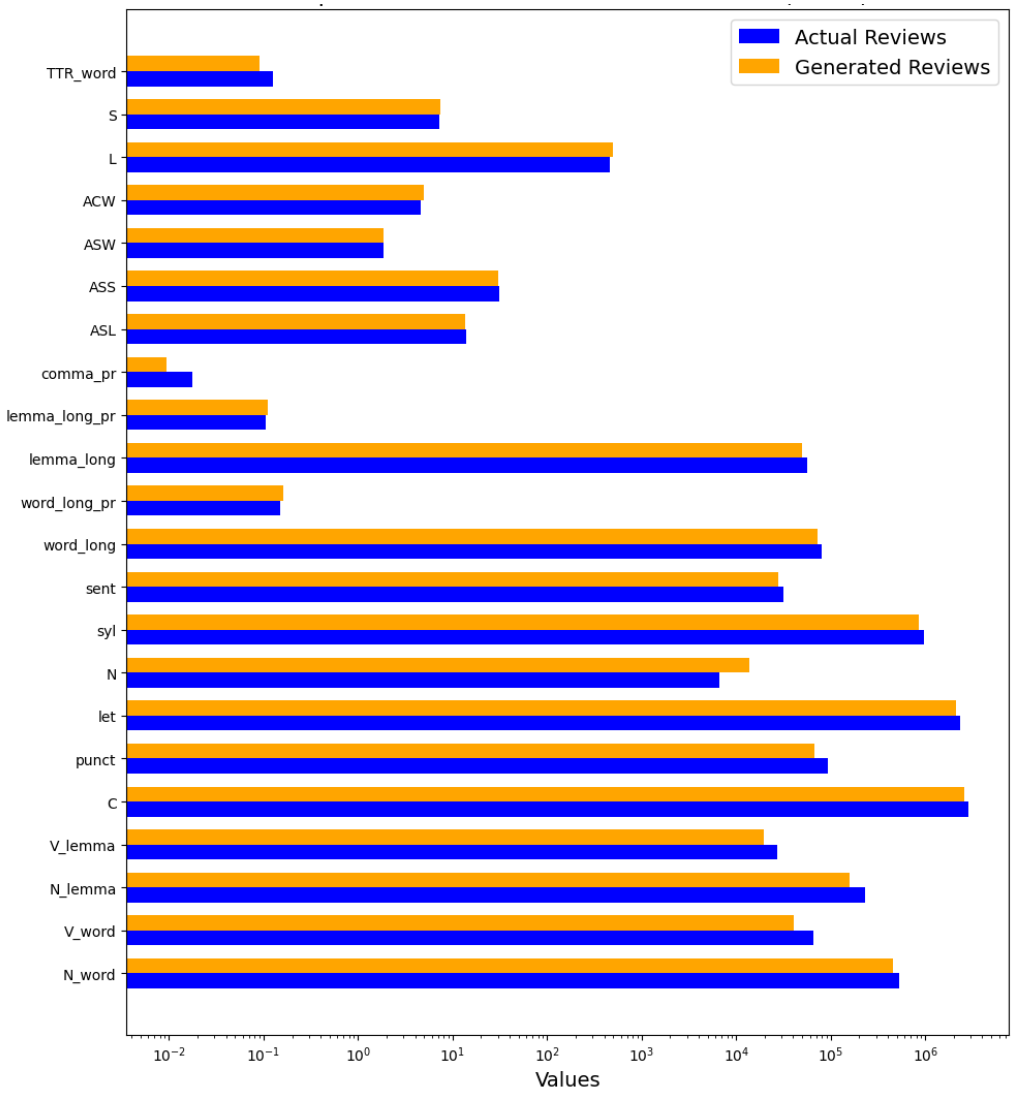
5.2. Text complexity assessment

To compute the values of a set of 130 metrics of text complexity in two subcorpora, we partially rewrote and ran the code for the model previously elaborated in (Blinova, Tarasov, 2022). On Figures 3-7 below we demonstrate the obtained results. As the majority of metrics are similar, we comment only those which show differences.

The Comma proportion metric (comma_pr) testifies that generated texts have less commas than written texts and the Number of numeric characters (N) metric – that generated texts contain more numbers than written reviews (Figure 3).

Figure 3. Comparison of values of text complexity metrics in written reviews (actual) and generated reviews. Part 1

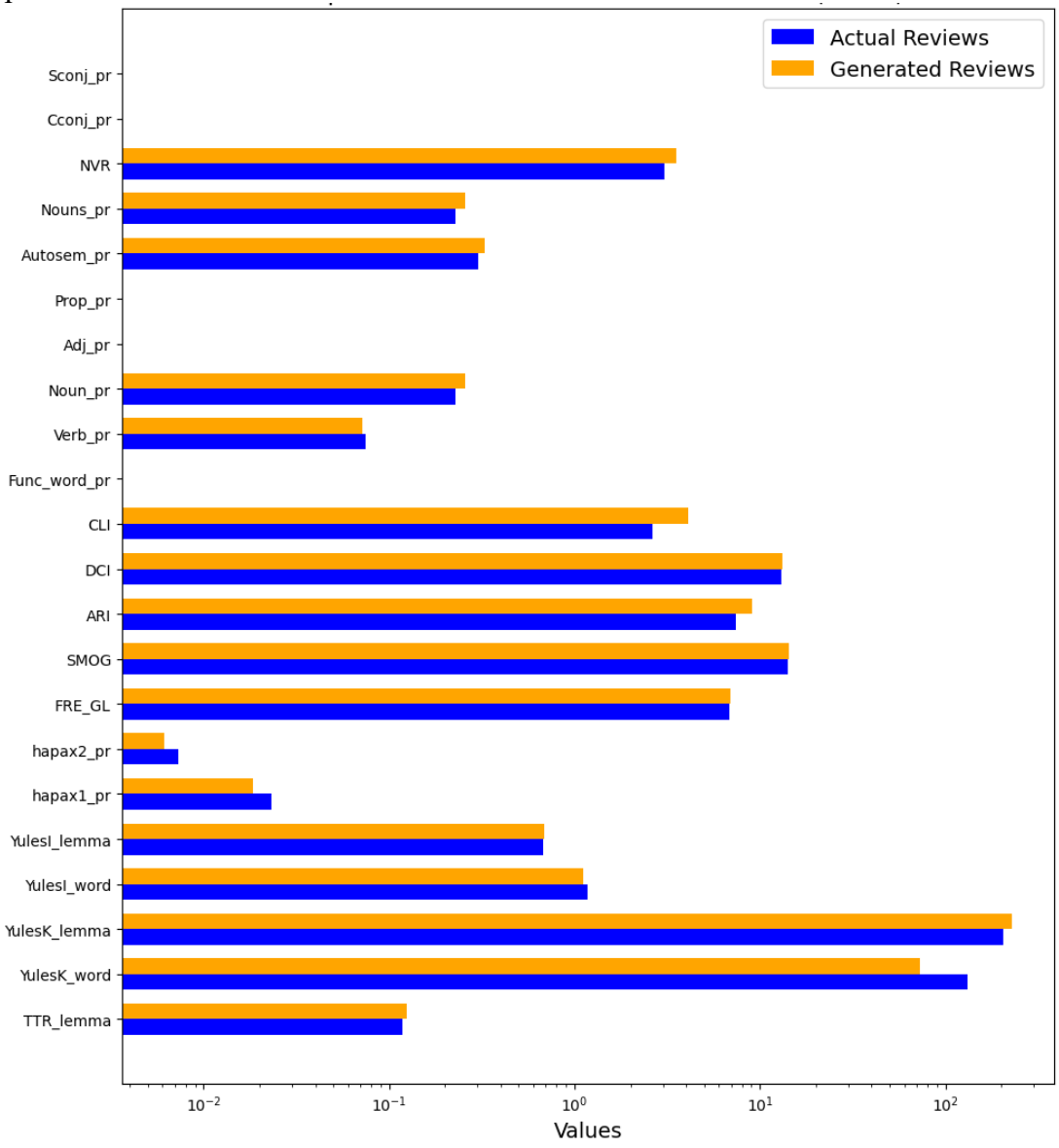
Рисунок 3. Сравнение значений метрик сложности текста в двух выборках: написанных и сгенерированных отзывах. Часть 1



For two readability indices (Coleman-Liau index (CLI) and ARI) out of five (Figure 4), the metric values are higher for generated texts (i.e., they are more difficult to understand) than for written texts. When comparing the values of the other three metrics, no differences were found (Flesch reading easy (FRE_GL), SMOG and Dale–Chall Index (DCI)).

Generated reviews also have lower values for Proportion of hapax legomena (for lemmas) and Proportion of hapax dislegomena (for lemmas) (hapax1_pr and hapax2_pr) than written reviews. This means that when generating a text, models are more likely than humans to repeat words they have already used in that text (rather than new occurrences). This is also indicated by the lower values of the Yules K metric for words (YulesK_word), which is commonly used to assess the lexical richness of a text.

Figure 4. Comparison of values of text complexity metrics in written reviews (actual) and generated reviews. Part 2
Рисунок 4. Сравнение значений метрик сложности текста в двух выборках: написанных и сгенерированных отзывах. Часть 2

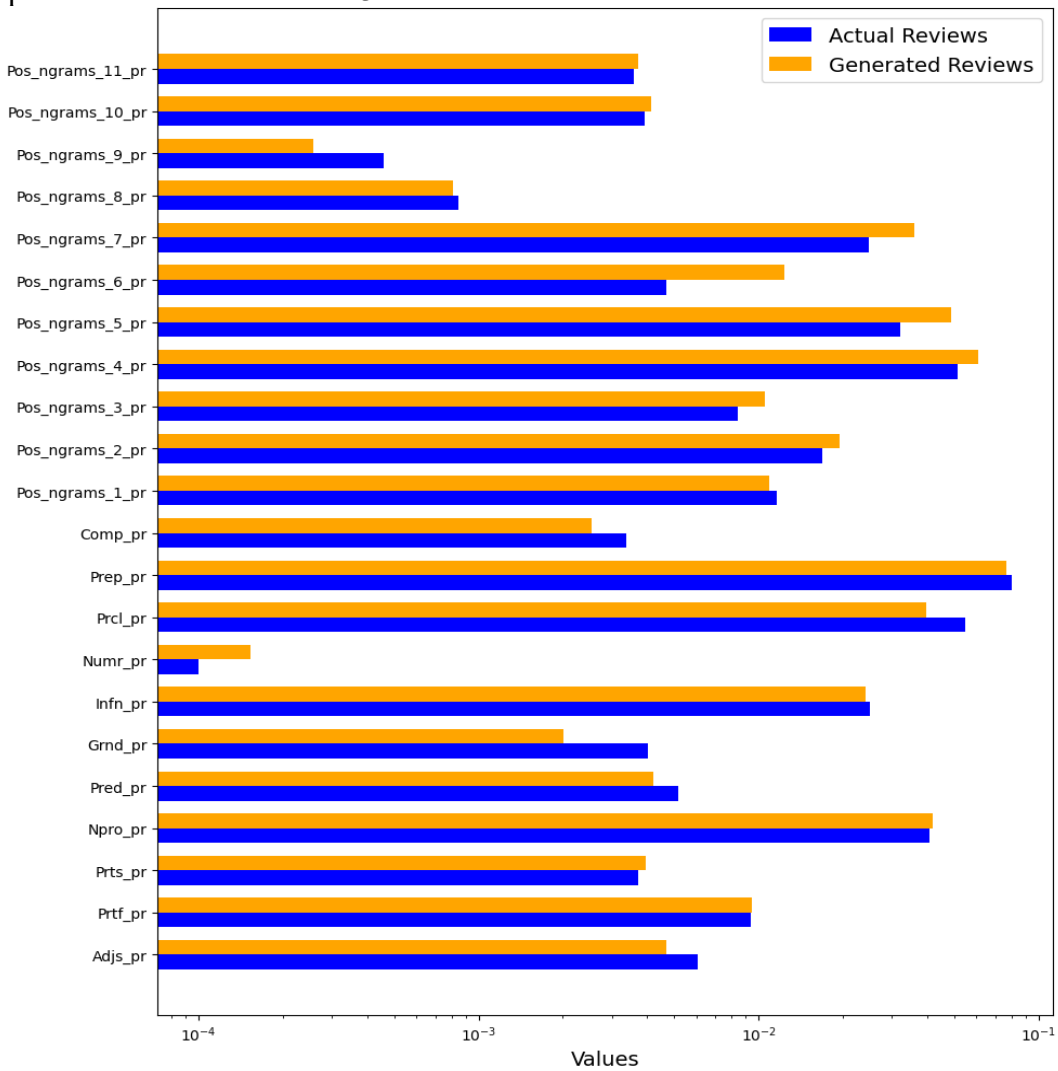


Metrics measuring the proportions of stable part-of-speech patterns show that, compared to written reviews, the generated texts contain significantly more bi- and trigrams, including nouns (Figure 5): NOUN + VERB (Pos_ngrams_2_pr), ADJF + NOUN (Pos_ngrams_4_pr), NOUN + NOUN (Pos_ngrams_5_pr), NOUN + NOUN + NOUN (Pos_ngrams_6_pr).

However, generated texts are inferior to written texts in terms of the number of the part of speech bigram "ADVERB+GERUND" (Figure 5 – Pos_ngrams_9_pr). Likewise, generated reviews contain fewer adverbial participles (Grnd_pr), short adjectives (Adjs_pr), comparative forms (Comp_pr) and predicatives (Pred_pr) than written texts. The latter are inferior to the generated texts in terms of the number of numeric characters (Proportion of numerals (Numr_pr)) (Figure 5).

Figure 5. Comparison of values of text complexity metrics in written reviews (actual) and generated reviews. Part 3

Рисунок 5. Сравнение значений метрик сложности текста в двух выборках: написанных и сгенерированных отзывах. Часть 3



Regarding the distribution of case proportions (metrics Ablt_pr and Gen_pr), human-written texts have slightly fewer noun forms in the Ablative and Genitive than generated texts, but more in the Locative (loct, Figure 6). If we examine the distributions of words of different ranks according to Zipf's law we surprisingly find that generated reviews contain more words of rank 1 and 2 (the less frequent ones) than written reviews.

Taking into account the genre type of texts from our collection, we explain this observation by the fact that LLMs are more attentive to the details of film plots or personage names, professions etc., than humans are. For the other ranks, there are practically no differences.

Figure 6. Comparison of values of text complexity metrics in written reviews (actual) and generated reviews. Part 4
Рисунок 6. Сравнение значений метрик сложности текста в двух выборках: написанных и сгенерированных отзывах. Часть 4

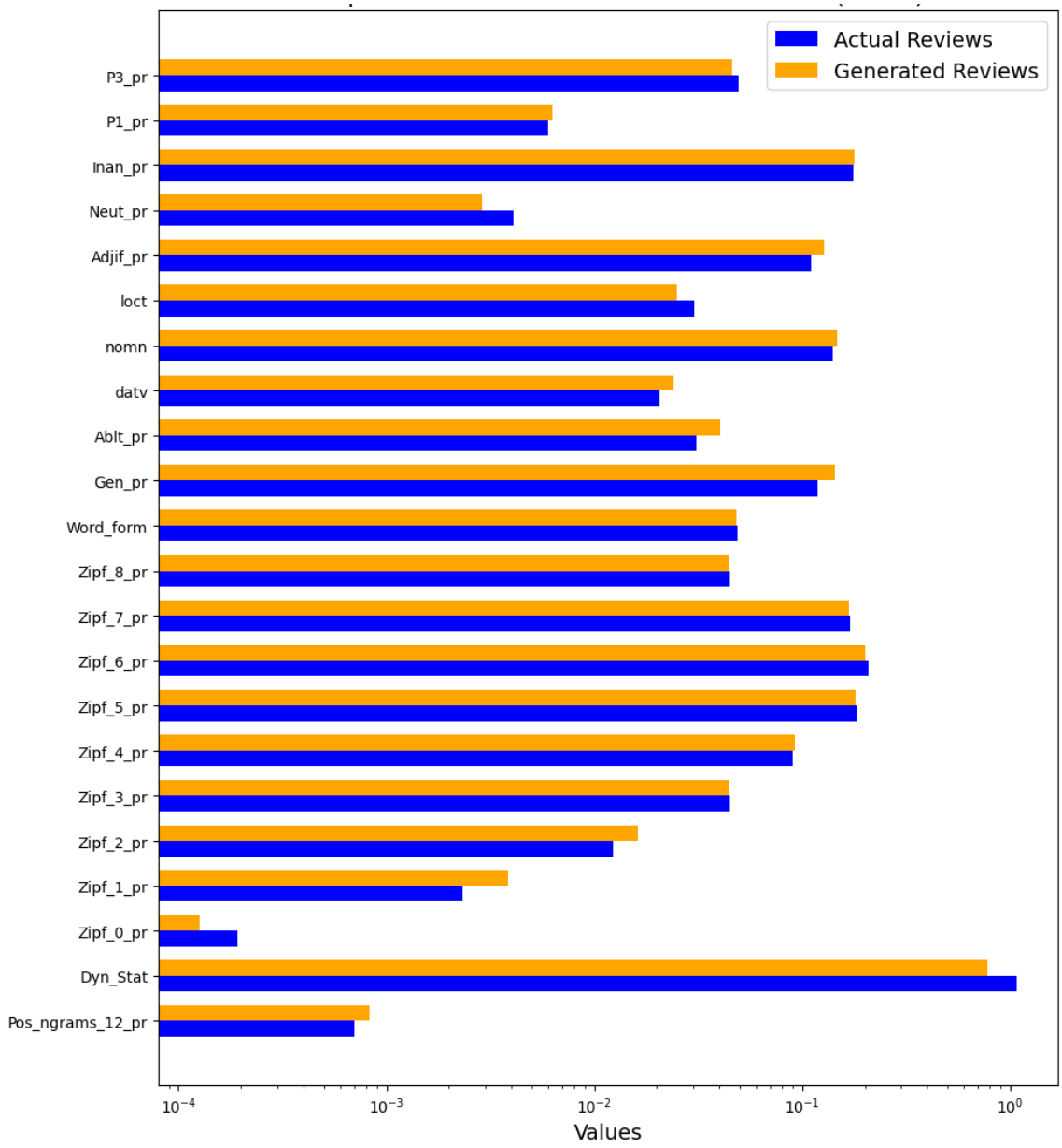
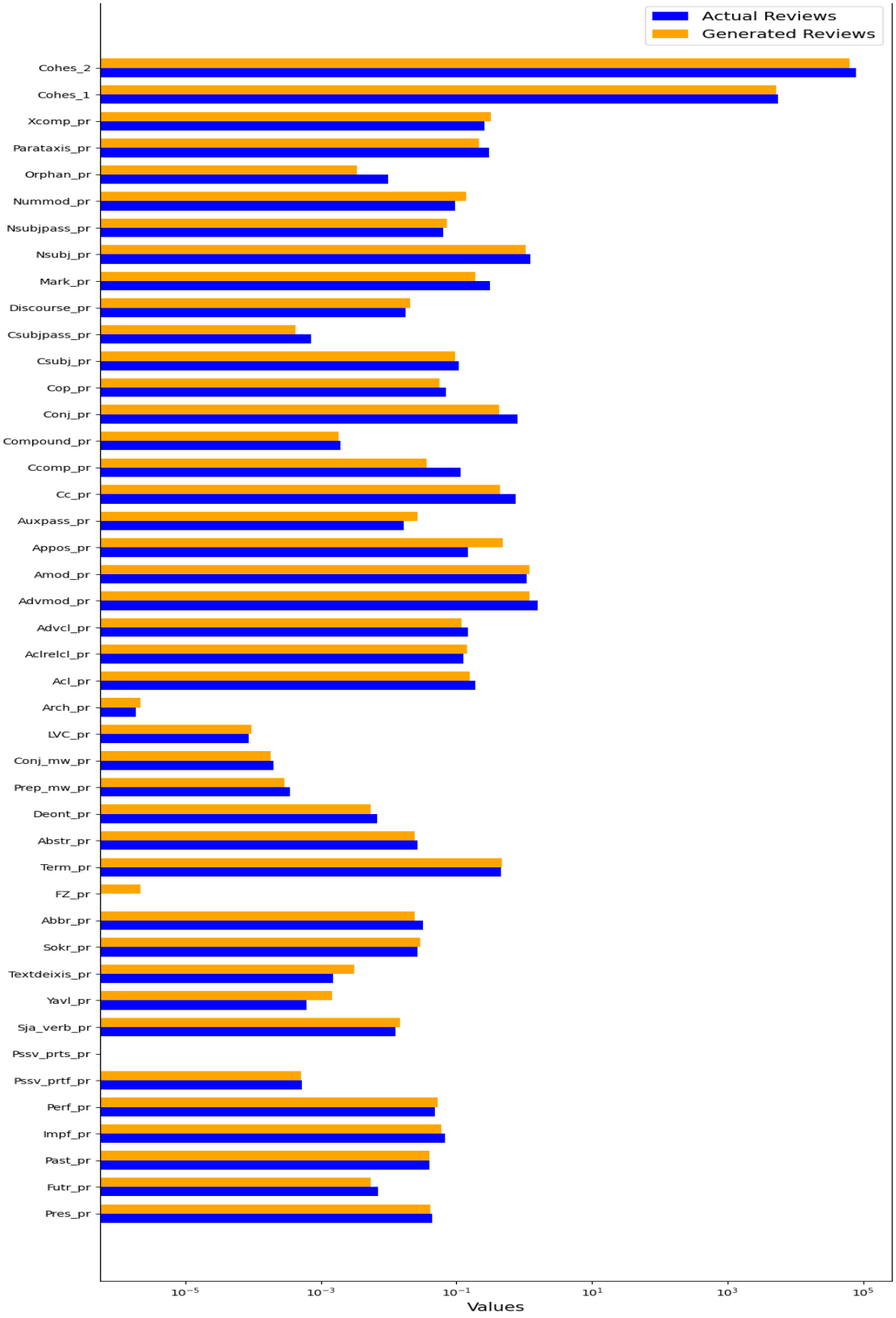


Figure 7. Comparison of values of text complexity metrics in written reviews (actual) and generated reviews. Part 5
Рисунок 7. Сравнение значений метрик сложности текста в двух выборках: написанных и сгенерированных отзывах. Часть 5



As for syntactical metrics (Figure 7), they testify that in written texts syntactical structures seem to be slightly more “fancy” than in reviews generated by LLM. For example, such metrics as the Proportion of elliptical predicate constructions (Orphan_pr), Proportion of constructions with clausal complements (Ccomp_pr, ex.g: *Мне кажется* (clausal complements)*, фильм был снят наспех; Он убедил меня в том, что фильм хорош* (clausal complements)), Proportion of conjunctions associated with conjuncts by the syntactic relation "cc" (coordination) (Cc_pr, ex.g: …*как детям, так и взрослым*), share of passive constructions "with a sentential subject" (Csubj:pass_pr, ex.g.: *Поднималась проблема о необходимости жертвовать ради других*) have higher values in written texts than in generated. For instance, compare fragments of two reviews from our parallel dataset (Table 4): in the written text we find four cases of elliptical predicate constructions, in the generated – only one; in the written – three cases of clausal complements, in generated – only one.

Table 4. Two texts from the parallel dataset to demonstrate difference in elliptical predicate constructions (EPC) and clausal complements (CC)
Таблица 4. Два текста из параллельного датасета: отличия в использовании эллиптических предикативных конструкций и сентенциональных дополнений

Written review	**Text generated by prompt from the written review**
Юриспруденция – тонкая наука. **Вопрос правосудия – гибкий вопрос (**EPC**). Вопрос коррупции (**EPC). Вот так автор сценария Курт Уиммер, приложивший руку до того к «Афере Томаса Крауна» и «Ультрафиолету», свёл воедино избитый сюжет триллера с драматизмом жестокой и интеллектуальной мести. В фильме режиссёра Ф. Гэри Грея дело не в том, **кто действительно виновен (**CC), а в том, **что «крутой» помощник окружного прокурора сможет доказать (**CC). Но система на то и система, **что с ней обязательно вступят в конфликт (**CC). **В работе Ника Райса (Джейми Фокс) короткая память – лучший помощник** (EPC)**. Главное – понравиться присяжным...** (EPC) [Jurisprudence is a subtle science. The issue of justice is a flexible issue. The issue of corruption. That’s how the screenwriter Kurt Wimmer, who had previously had a hand in "The Thomas Crown Scam" and "Ultraviolet", brought together the hackneyed plot of the thriller with the drama of cruel and intellectual revenge. In the film directed by F. Gary Gray, the point is not who is really guilty, but what the "cool" assistant district attorney will be able to prove. But the system is also a system that will necessarily come into conflict with it. In the work of Nick Rice (Jamie Foxx), short memory is the best assistant. The main thing is to please the jury...]	Юриспруденция – тонкая наука. Вопрос правосудия в ней не стоит так остро, **как это может показаться на первый взгляд (**CC). Но именно эта тонкость и делает ее столь важной областью знаний для каждого из нас. **В фильме «Грязь» Брюс Робертсон (Джеймс МакЭвой) – шотландский полицейский, расследующий серию жестоких убийств молодых девушек (**EPC). Для расследования этих преступлений он использует самые изощренные методы психологического давления, запугивания и шантажа. И все бы ничего, если бы дело не выеденного яйца не стоило… [Jurisprudence is a subtle science. The issue of justice in it is not as acute as it may seem at first glance. But it is precisely this subtlety that makes it such an important area of knowledge for each of us. In the film Dirt, Bruce Robertson (James McAvoy) is a Scottish policeman investigating a series of brutal murders of young girls. To investigate these crimes, he uses the most sophisticated methods of psychological pressure, intimidation and blackmail. And it would have been all right if it hadn’t been worth a damn.…]

However, generated texts of review contain more appositional modifiers (Proportion of appositional modifiers − Appos_pr, ex.g.: *Элизабет, любовница главного героя, принимает решение*...), deictic words (Textdeixis_pr, ex.g.: *этот фильм вышел в 2017 году*) and forms of the verb *являться* (equivalent of *to be* in English; Yavl_pr) than texts produced by humans.

Summarizing the results, we can say that, although in general the differences in the values of the metrics cannot be called significant, several trends can be seen.

Firstly, written texts, unlike generated texts, have a slightly more complex syntactic structure (more commas, ellipses, clausal complements, etc.). These syntactic features are induced by the desire of the speaker, a social subject, to express their personal judgment, opinion.

Second, it appears that human authors are more likely to vary lexical units and constructions than generative models. The latter, on the other hand, show a predilection for part of speech bi- and trigrams including a noun or even a chain of nouns. They also avoid non-nucleus forms of expressing predicativity and are predisposed to use the vocabulary of low and medium frequency.

Thirdly, the generated texts, if compared to written reviews, contain a greater number of numerical symbols and deictics, which is probably due to the greater attention to details seeable in the synthetical reviews.

5.3. Human evaluation test

In our human evaluation test, we obtained a total of 4534 evaluations of 36 mixed (synthetical and human-written) texts from 130 informants. As some informants did not evaluate the entire sample, the number of estimates collected is less than expected.

Only 25% of synthetical texts from our mixed sample were identified correctly; in 75% of responses for this group, generated texts were taken for human-written.

As for the human-written texts, in 39% of responses the written texts from our mixed sample were wrongly labeled as synthetical; in 61% of cases label was attributed correctly (Table 5).

Table 5. Results of assessment
Таблица 5. Результаты оценки

Type of text	% of true positives	% of false positives
Generated	25 in corpus of assessments for generated texts	75 in corpus of assessments for generated texts
Written	61 in corpus of assessments for written texts	39 in corpus of assessments for written texts

It suggests that participants found it more difficult to differentiate artificial reviews generated by LLM from human-written reviews and less difficult to differentiate human-written reviews from generated ones.

After having rated all texts, the informants were asked to formulate some principals they were guided by while deciding about the naturalness and human-written or AI generated character of texts they were exposed to. In Table 6 we summarize our informants' intuitions.

Table 6. Synthetical texts and human-written texts features as they are assumed by informants according to their comments ("+" marks the features that are thought being proper to the text category, "-"– not proper)
Таблица 6. Признаки сгенерированных и написанных (естественных) текстов по интроспективным ощущениям информантов, отраженным в комментариях ("+" – признак присутствует, "-" – признак присутствует)

	Feature	Synthetical	Human-Written
	Syntax		
1	Parcellations	+	-
2	Long enumerations	+	-
3	Sentences formed like definitions – X it is... (*Одиночество – это когда ты никому не принадлежишь, но при этом чувствуешь себя несчастным из-за отсутствия близкого человека рядом с тобой [Loneliness is when you don't belong to anyone, yet you feel miserable because of the lack of a loved one by your side])*	+	-
4	Long and complex sentences (with coordination and/ or subordination) with many dots by the end (e.g. *Также стоит отметить игру Брэда Питта и Хелены Бонэм Картер. Их персонажи получились весьма колоритными и запоминающимися [Also worth noting is the performance of Brad Pitt and Helena Bonham Carter. Their characters turned out to be very colorful and memorable]*)	-	+
5	Non motivated word repetitions (e.g. *Эмили Блант весь фильм ходит с одним и тем же выражением лица на протяжении всего фильма [Emily Blunt walks around the entire movie with the same facial expression throughout the entire movie]*)	-	+
	Lexical variability		
6	Very typical phrases and clichéd turns (e.g. *Если вы хотите посмотреть легкое кино, чтобы отвлечься вечером после трудового дня…* [*If you want to watch a light movie for an evening viewing after a hard day's work…*])	+	-
7	Use of colloquialisms, jargons and verbal markers of hesitation	-	+
	Stylistic featuring		
8	Absence of complex stylistic devices based on semantic mechanisms (litotes, metaphors, oxymora)	+	-
9	Irony, linguistic game based on polysemy	-	+
10	Wrong collocations, misuse of lexical items	+	-
11	Specific "strange" style inherent to the entire text (ex.eg. *you look into your soul and try to see yourself differently, to understand that you have nothing in common with Them – and realize that you are also one of Them, you are a beast*)	-	+
	Textual categories		
12	Lack of logical coherence (ex.g. *Я не могу объяснить почему так вышло – может быть это из-за того что фильм слишком детский? Или же он мне показался чересчур жестоким, но все равно после просмотра осталось ощущение пустоты и грусти*	+	-

	[I can't explain why it was like that – maybe it was because the movie was too childish? Or maybe I found it too violent, but still I was left with a feeling of emptiness and sadness after watching it.])		
13	Intertextual and intergenres insertions (*quotations, forms proper to another speech genre*)	-	+
14	Too much of logical coherence, repetitive structures	+	-
15	References to personal experience	-	+
16	The repetitiveness of text composition (plot, critical opinion, compliments to the actors, camera men etc.)	+	-
17	Variability in text composition	-	+
18	Factual errors	+	-
19	The abundance of details and facts	+	-
20	Predominance of attitudes and emotions over facts and details (*И все-таки мне непонятно, почему эта картина получила столь высокую оценку у критиков и завоевала столько положительных отзывов зрителей?! [Still, it's unclear to me why this film was so highly praised by critics and won so many positive reviews from viewers!*])	-	+
	Text affectivity		
21	Unmotivated changes in text sentiment (e.g. Что происходит? Что они хотели сказать этим фильмом? Я так и не нашел ответ на мои вопросы. Подбор актеров прикольный [*What's going on? What did they want to say with this movie? I never found the answers to my questions. The cast is amazing!]*)	+	-
22	Explicitly manifested, mostly positive, attitude towards film creators and participants, but mostly negative − towards film critics	-	+
23	General positive sentiment *(Фильм гениален, игра актеров великолепна. Прекрасный фильм [The film is brilliant, the acting is superb. It's a wonderful film])*	+	-
24	Emoji	-	+
25	Emotional waves (*Да он хорош в плане графики (особенно сцены с самолётами), да это хороший сюжетный ход для фильмов о будущем (хотя я сомневаюсь что будущее будет хорошим) Но почему то всё остальное хромает [Yes, it's good in terms of graphics (especially the aeroplane scenes), yes it's a good storyline for a film about the future (although I doubt the future will be good) but for some reason everything else is lame]*).	-	+

Thus, in this table (Table 6), we have tried to summarize how informants conceive "naturalness" (i.e., characteristics of exclusively "human" use of natural language for the task of writing a review) and "artificiality" (i.e., characteristics inherent in the way artificial intelligence uses natural language for the same task).

Naturalness is conceptualized by informants as a set of text characteristics that are a projection of informants' stereotyped representation of human thinking abilities, such as (numerals in brackets are assigned according to the linguistic features listed in Table 6):

- the human ability to formulate complex and multi-component judgements (4);
- the possible aberrations in human thought processes (5);

- emotionality of thinking and the tendency to lose rational control in favor of emotionality (20, 24);
- mind flexibility permitting to the Humans to merge entity by sophisticatedly combining heterogeneous elements (9, 13, 17);
- mind rootedness in personal life and social routines (15, 22).

As for unnaturalness representation (syntheticalness), the informants conceptualized it as a projection of their stereotypes about AI thinking abilities, such as:

- stereotypical flow of thoughts (2; 6);
- limited set of cognitive structures and their repetitiveness (14, 16);
- categorization based on only explicit entities features, incapability to deal with implicit links (3);
- erratic patterns of thinking and memorizing things (12, 18);
- exactitude of memorizing items even erratic (19);
- no link with life experience or social routines (21, 23).

We supposed that this kind of viewing could be partially conditioned by autostereotypes of humans about themselves and their heterostereotypes about "machines" and as such could be misleading in defining what text is more natural.

To verify this, we did an expert analysis of a sample of mislabeled texts: written texts wrongly assessed by our group of informants as being synthetical or synthetical texts wrongly assessed as written. Our main assumption was:

- we know the true label of the texts;
- informants wrongly placed them in the category which is not appropriate;
- if we find in written texts a considerable number of features of synthetical texts / in generated texts a number of features of written texts as they are perceived by informants, then human evaluators are misled by their intuitions and then – they actually trust them.

Below, we consider two examples of **written texts wrongly judged by assessors to be synthetic**.

Text 1 (false positive: written assessed as generated)

57 % of informants categorized it as synthetic, 43% – human-written

Этот фильм ужасен, ужасен в том смысле, что он показывает человека с такой стороны, с которой никто бы не хотел его видеть. Это фильм о простых людях, которые получили власть, пусть маленькую, над одним человеком, но все же власть. В этом шедевре мирового кинематографа показывается, как все-таки могут люди ненавидеть друг друга и в кого они могут превращаться: в псов, в животных или еще ниже. Надо сказать, что это очень тяжелый фильм, но смотрится на одном дыхании от начала до конца, и ты не замечаешь, как проходит три часа. Бесспорно 10 из 10 (*This movie is horrible, horrible in the sense that it shows a side of man that no one would want to see. It is a movie about ordinary people who have been given power, albeit small, over one person, but power nonetheless. This masterpiece of world cinematography shows how people can hate each other and who they can turn into: dogs, animals or even lower. I must say that it is a very heavy movie, but it is watched in one breath from the beginning to the end, and you don't notice how three hours pass. Undoubtedly 10 out of 10*).

Table 7. Features of naturalness and artificialness from informants' comments (each number refers to the text feature given in Table 6) detected by experts in the Text 1
Таблица 7. Обнаруженные экспертами в Тексте 1 признаки естественности и искусственности, сформулированные информантами в комментариях (номера соответствуют порядковому номеру признака в таблице 6)

Synthetical texts' features in the Text 1	**Natural texts' features in the Text 1**
"В этом шедевре мирового кинематографа" [This masterpiece of world cinematography] − 6	Этот фильм ужасен, ужасен в том смысле, что он показывает человека с такой стороны, с которой никто бы не хотел его видеть [This movie is horrible, horrible in the sense that it shows a side of man that no one would want to see] – 4
в псов, в животных или еще ниже" [dogs, animals or even lower] − 2	ужасен в том смысле… [horrible in the sense that it shows a side of man that no one would want to see] – 9
"смотрится на одном дыхании от начала до конца, и ты не замечаешь, как проходит три часа" [it is watched in one breath from the beginning to the end, and you don't notice how three hours pass] – 6	В этом шедевре [In this masterpiece]; этот фильм ужасен [this movie is horrible]; это очень тяжелый фильм… [it is a very heavy movie] – 20
Repetitive text structure: assumption – precision – 14	−

In Text 1, the experts detected the following features of a synthetic text: clichéd expressions and collocations, enumerations, repetitive structure of proposition "general assumption, precision" (*power, but only over one person; heavy film, but it is watched in one breath*). Among the written text features the experts found: playing with different senses of a polysemantic words (*horrible in the sense of...*), complex syntax (the first sentence), evaluation and opinion prevail over facts and details (evaluative epithets: *very heavy movie*).

Text 2 (false positive: written assessed as generated)

68 % of responses − synthetic, 32% – human-written

Фильм, является гениальным. Из-за его простоты, точней отсутствия в картине массы декораций (стены домов, трава и т.п.) он заставляет сосредоточиться исключительно на людях. Весь фильм ты проживаешь не реальным миром, его красотой, а его героями, людьми, их эмоциями и чувствами, переживаниями и размышлениями. Ты с головой окунаешься в мир человека нуждающегося в помощи. Ощущаешь доброту, проявленную с опаской, а после и вовсе ушедшей из жизни данных людей. Волки… Они вгрызаются все глубже, вырывая с каждым разом все больший кусок мяса… Страх – именно он делает нас животными, но разве есть у нас сила сопротивляться ему? Ты смотришь в свою душу и пытаешься увидеть себя другим, понять, что у тебя нет ничего общего с Ними – и осознаешь, что ты тоже один из Них, ты зверь. Меняться никогда не поздно…. Просмотр этой картины позволяет понять, что Людей в этом мире почти нет, и они не появятся из неоткуда… Надо стать ими. Нельзя продавать Добро, лишь утешая свой страх. Человек создан для того, чтобы бороться. И в первую очередь с самим собой… 10 из 10 [The movie is brilliant. Because of its simplicity, or rather the absence of a lot of scenery (walls of houses, grass, etc.), it makes

you focus solely on people. The whole movie you live not in the real world, its beauty, but in its characters, people, their emotions and feelings, experiences and reflections. You plunge headlong into the world of a person in need of help. You feel the kindness, shown with fear, and afterwards and completely gone from the life of these people. Wolves... They bite deeper and deeper, tearing out a bigger piece of meat each time... Fear is what makes us animals, but do we have the power to resist it? You look into your soul and try to see yourself differently, to understand that you have nothing in common with Them – and realize that you are also one of Them, you are a beast. It's never too late to change.... Viewing this film makes you realize that there are almost no People in this world, and they will not appear out of nowhere... You have to become them. You can't sell the Good just to comfort your fear. Man was created to fight. And first of all with himself... 10 out of 10]

Table 8. Features of naturalness and artificialness from informants' comments (each number refers to the text feature given in the Table 6) detected by experts in the Text 2
Таблица 8. Обнаруженные экспертами в Тексте 2 признаки естественности и искусственности, сформулированные информантами в комментариях (номера соответствуют порядковому номеру признака в таблице 6)

Synthetical texts' features in the Text 2	**Natural texts' features in the Text 2**
"Весь фильм ты проживаешь не реальным миром, его красотой, а его героями …" [The whole movie you live not in the real world, its beauty, but in its characters] – 10	"Страх – именно он делает нас животными, но разве есть у нас сила сопротивляться ему?" [Fear is what makes us animals, but do we have the power to resist it?] − 20
"… реальным миром, его красотой, а его героями, людьми, их эмоциями и чувствами, переживаниями и размышлениями" […the real world, its beauty, but in its characters, people, their emotions and feelings, experiences and reflections] – 2	Specific "strange" style inherent to the entire text − 11
Ощущаешь доброту, проявленную с опаской, а после и вовсе ушедшей из жизни данных людей. Волки… Они вгрызаются все глубже, вырывая с каждым разом все больший кусок мяса… [You feel the kindness, shown with fear, and afterwards and completely gone from the life of these people. Wolves... They bite deeper and deeper, tearing out a bigger piece of meat each time...]– 12	Из-за его простоты, точней отсутствия в картине массы декораций [Because of its simplicity, or rather the absence of a lot of scenery] − 9
Просмотр этой картины позволяет понять, что Людей в этом мире почти нет, и они не появятся из неоткуда… Надо стать ими [Viewing this picture makes you realize that there are almost no People in this world, and they will not appear out of nowhere... You have to become them] – 1	–

The analysis demonstrates that a number of characteristics of the text under discussion (written text) do indeed bring it closer to the generated text as it is perceived by informants: violation of lexical combinatorics (*Весь фильм ты проживаешь не реальным миром, его красотой… – проживать*+Ablativus), enumeration (*его героями, людьми, их эмоциями…*), broken logical connection (a passage that first talks about people and their experiences and then suddenly conjures up the image of wolves), and parcellations (*Людей в этом мире почти нет, и они не появятся из ниоткуда… Надо стать ими*).

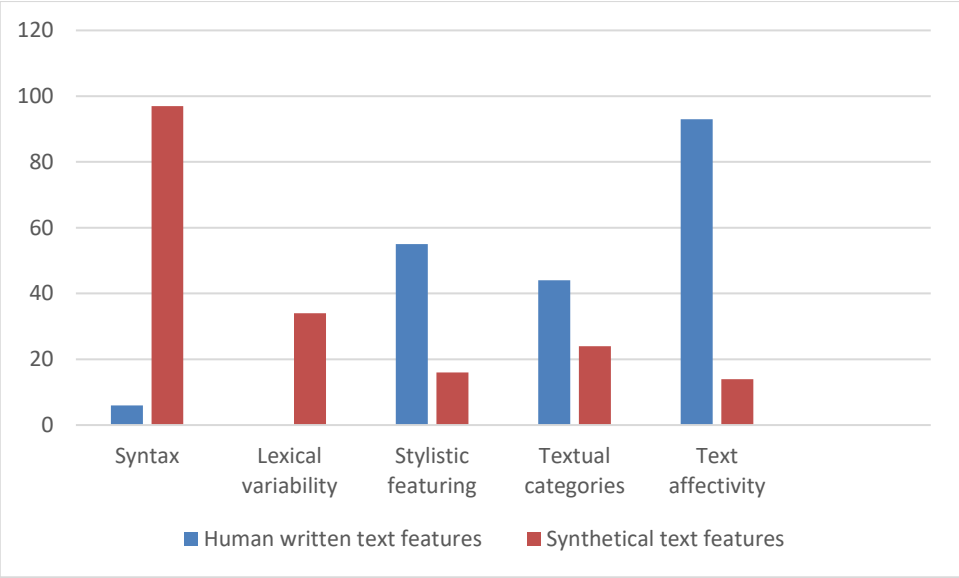
At the same time, the text fragment contains some human specific features: a sort of semantic play based on the polysemy of the noun *простота* (1. simplicity; 2. Naiveness: *Из за его простоты, точней отсутствия в картине массы декораций*); rhetoric question, metaphors which testify the predominance of attitudes and emotions over facts and details (*Страх – именно он делает нас животными, но разве есть у нас сила сопротивляться ему?*); specific "strange" style inherent to the entire text.

As we can see, when assessing the naturalness of texts, our informants overestimate the human ability to formulate complex judgments and the flexibility of the mind, and at the same time – the stereotypical way of thinking typical of AI. As we see, actually, humans could produce texts by altering both strategies: repetitive patterns and those which are based on their surprising "fuzzy" logic, motivated by lived life experience and imagination.

When we applied the same analytic procedure of feature annotation to the subsample of synthetic texts labeled as written, we obtained the following distribution of features in them (Figure 8). The bar charts show that the informants, while assessing texts, were actually misled by features connected to such categories as text affectivity, stylistic devices (e.g. "comparison" *Режиссер словно пытается переписать на модный музыкальный центр старую затасканную кассету* [It's like the director is trying to re-record an old hackneyed cassette tape on a fancy music centre]) and textual categories (e.g. "reference to personal experience" *Посмотрел этот фильм по совету друзей* [Watched this film on the advice of friends]). Being charmed by Human likness of emotions, style and text "actualness" the informants didn't focus on wrong lexical combinations (e.g. *Но нет намёка* ***о том***…) nor syntactical trivialness (e.g. *Великолепная актерская игра, тонкий юмор, потрясающий саундтрек делают этот фильм непревзойденным* [Great acting, subtle humour, and a terrific soundtrack make this film second to none]).

In other words, those informants who focused their attention on textual, stylistic and affective categories of text characteristics, overestimating their importance due to the influence of the mentioned above autostereotypes, made the wrong decision about text attribution.

Figure 8. Distribution of features in the sample of synthetical texts wrongly attributed by informants to human-written
Рисунок 8. Распределение признаков в корпусе сгенерированных текстов, ошибочно атрибутированных информантами к категории написанных человеком



6. Discussion

Naturalness in the "pre-AI" era was viewed from the dichotomy of "an experienced language user within a given community vs a person who has not fully mastered language community's routines". In the AI era, the scale of naturalness has acquired another vector: who is the author of the text - a human (in which case the text is natural) or a "machine" (in which case it is artificial or synthetic)?

Our study attempted to look at naturalness within this latter scale from three perspectives: a formal metric for assessing the differences between two adjacent sentences (self-BLEU), metrics of linguistic text complexity, and a psycholinguistic experiment.

The values of self-BLEU metric measured for human-written texts showed that naturalness at this side can be regarded as a relatively low level of difference between two adjacent sentences. As linguists, we understand that this is due to the realization of the category of cohesion (grammatical, lexical) as an integral category of every normal human-written text. However, this formal approach does not seem sufficient for defining naturalness.

From very linguistic point of view, **naturalness could be seen as a text characteristic that is due to human ability to get access to first-order mental structure** (Thibault, 2011) **(e.g. "mental spaces" in Fauconnier's sense** (Fauconnier, 1981)**, or "scene" – in Talmy's theory** (Talmy, 2000)**) using second-order structures − words and grammatical patterns**. While processing words, language users keep the whole imagined scene in their cognitive view. This enables specific linguistic properties of "natural" text:

- Predicativity that dominates nominativity;
- Sentence syntactical incompleteness that does not interfere with understanding;
- Sentences with sentential subject and complement;
- Comparative constructions;
- Mid- and high frequency vocabulary, but diverse.

Since the speaker always has access to the whole first-order mental structure (scene

representation), they can easily divide it into different segments and then – by means of language (second-order structures) – establish complex and multilevel relations between them (such as causality; anteriority – simultaneity – posteriority; comparison; second-order structures, etc.). The language users feel free to skip some parts of second-order structures (e. g. ellipses), shorten them, because they are easily reconstructed by the hearer/ reader, who, while understanding, tries to build up an analogously accessible representation (first-order structure) of the described scene. These iteractions between second-order and first-order structures make it possible to name entities of mental structures by using diverse words according to the movement of the window of cognitive attention. This capacity to have simultaneous access to the structures belonging to different cognitive orders while writing or speaking differs Humans from AI because the latter generates texts by operating on the same order of structures (tokens, lemmas) using the principle of "guess-the-next-word". According to Bender et al., 2021, "LM is a system for haphazardly stitching together sequences of linguistic forms it has observed in its vast training data, according to probabilistic information about how they combine, but without any reference to meaning: a stochastic parrot" (Ibid: 617).

In other words, when prompting or training a model we should encourage the LLM to pay attention to the mentioned above features to gain in naturalness.

However, in psycholinguistic perspective, **the naturalness appears as a set of expectations that people have from themselves as Humans**. When assessing texts and deciding who is the author (Human or AI), people anticipate these expectations overestimating text features that correspond to them. Psycholinguistically, **naturalness is conceptualized as a collection of text features that reveal human ability:**

- **to formulate complex judgments;**
- **to merge very different entities on the basis of fuzzy logic principles;**
- **to be emotion dominated;**
- **to be socially sensitive, sometimes affected by logic aberration;**
- **to be able to give reference to his personal life**.

If we want to generate text with a maximum of human likeness, we should also take these pertinent expectations into account when collecting training data or thinking about prompting strategies (Wei et al., 2023).

Obviously, there is some correlation between the concept of naturalness as it is seen in those two perspectives: linguistic and psycholinguistic. For example, the informants' expectation that Humans produce more complex thoughts than AI finds its counterpart in such linguistic features of human texts as high values of the metrics of Proportion of elliptical predicate constructions, Proportion of clausal complements, Proportion of units capable of attaching dependent clauses etc. Nevertheless, there are many expectations manifested by the informants in experimental study, but still hardly detectable by text complexity metrics (e.g. social sensitiveness or fuzzy logic principles in combining thoughts and words).

7. Conclusion

The conducted research has shown that in the situation of the permanently growing power of LLMs, the category of text naturalness needs to be revised. Nowadays, when thinking about naturalness, we mean such text properties which are "inalienable" from Humans, properties which are derived from human nature in itself and cannot be imitated by AI.

To define text naturalness in such context, we created a parallel corpora of human-written film reviews and reviews generated by using prompts from the written reviews; to compare corpora, we applied a formal metric of text diversity, a set of metrics of text complexity and we organized a psycholinguistic experiment by inviting informants to assess text naturalness, to label each text "written" or "generated", and to articulate their introspective intuitions about text naturalness features.

Our comparative analysis of text complexity metrics in written and generated texts didn't demonstrate any absolutely significant differences in metrics values. However, by summarizing the majority of discrepancies existing between values obtained in two compared subcorpora, we gave a definition of naturalness as a text category: it is **a set of text features allowed by human ability to simultaneously operate by first-order structures (mental representations) and second-order structures (words and grammatical constructions)**. In practice, such naturalness manifests itself in more complex and nuanced syntactic relations between sentence parts, possibility to vary the level of syntactic structures completeness, rather high word diversity, use of sophisticated punctuation.

Taken as a psycholinguistic category, naturalness can be perceived as **a set of intersubjective expectations based on autostereotypes that Humans privilege while thinking about themselves: some of them are true, other – illusory.** In practice, such naturalness manifests itself in linguistic means to express emotions, to appeal to life experience, to avoid all kinds of repetitions, to generate new meanings by using stylistic devices and implicit links.

The obtained results also showed some new perspectives in detecting naturalness. The first interesting thing to do in further research is to compare how Zipf's law is fulfilled in natural and syntactic texts. The second consists in training LLM according to informants' intuitions collected in our experiment and to do another assessment.

References

Alzahrani, E. and Jololian, L. (2021). How Different Text-Preprocessing Techniques Using The BERT Model Affect The Gender Profiling of Authors, *arXiv preprint arXiv: 2109.13890.* https://doi.org/10.48550/arXiv.2109.13890 (*In English*)

Bally, Ch. (1913). *Le language et la vie*, Edition Atar, Paris, France. (*In French*)

Bender, E. M., Gebru, T., McMillan-Major, A. and Shmitchell, Sh. (2021). On the dangers of stochastic parrots: Can language models be too big?, *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. (*In English*)

Blinova, O. and Tarasov, N. (2022). A hybrid model of complexity estimation: Evidence from Russian legal texts, *Frontiers in Artificial Intelligence*, 5. https://doi.org/10.3389/frai.2022.1008530 (In English)

Celikyilmaz, A., Clark, E. and Gao, J. (2021). Evaluation of text generation: A survey, *arXiv preprint arXiv*: *2006.14799.* https://doi.org/10.48550/arXiv.2006.14799 (*In English*)

Dashela, T. and Mustika, Y. (2021). An Analysis of Cohesion and Coherence in Written Text of Line Today about Wedding Kahiyang Ayu and Bobby Nasution, *SALEE: Study of Applied Linguistics and English Education*, 2 (2), 192−203. https://doi.org/10.35961/salee.v2i02.282 (*In English*)

Fauconnier, G. (1981). Pragmatic functions and mental spaces, *Cognition*, 10 (1-3), 85−88. (*In English*)

Holtzman, A., Buys, J., Du, L., Forbes, M. and Choi, Y. (2019). The Curious Case of Neural Text Degeneration, *arXiv preprint arXiv: 1904.09751.* https://doi.org/10.48550/arXiv.1904.09751 (*In English*)

Lavie, A. &Agarwal, A. (2007). METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments, *Proceedings of the Second Workshop on Statistical Machine Translation*, 228–231. (*In English*).

Li, C., Zhang, M. and He, Y. (2022). The Stability-Efficiency Dilemma: Investigating Sequence Length Warmup for Training GPT Models, *arXiv preprint arXiv*: *2108.06084v4.* https://doi.org/10.48550/arXiv.2108.06084 (*In English*)

Lin, Ch-Y. (2004). Rouge: A package for automatic evaluation of summaries, *Text summarization branches out*, 74–81. (*In English*)

Liu, X., Ji, K., Fu, Y., Lam Tam, W., Du, Zh., Yang, Zh. and Tang, J. (2022). P-Tuning v2: Prompt Tuning Can Be Comparable to Fine-Tuning Universally Across Scales and Tasks, *arXiv preprint arXiv: 2110.07602.* https://doi.org/10.48550/arXiv.2110.07602 (*In English*)

Margolina, A.V. (2022). Controlling impression: making ruGPT3 generate sentiment-driven movie reviews, *Journal of Applied Linguistics and Lexicography*, Vol. 4., 1, 15-25. (*In English*)

Margolina, A., Kolmogorova, A. (2023). Exploring evaluation techniques in controlled text generation: a comparative study of semantics and sentiment in ruGPTLarge-generated and human-written movie reviews, *Computational Linguistics and Intellectual Technologies*: Papers from the Annual International Conference, 1082-1090. (*In English*).

Mikhaylovskiy, N. (2023). Long story generation challenge, *Proceedings of the 16th International Natural Language Generation Conference: Generation Challenges*, 10–16. (*In English*)

Mnih, V., Kavukcuoglu, K., Silver, D. et al. (2015). Human-level control through deep reinforcement learning, *Nature*, 518 (7540), 529–533. http://dx.doi.org/10.1038/nature14236 (*In English*)

Newmark, P. (1987). *Manual de traducción*. Madrid: Ediciones Cátedra. (*In Spanish*)

Novikova, J., Lemon, O. and Reiser, V. (2016). Crowd-sourcing NLG data: Pictures elicit better data, *Proceedings of 9th International Natural Language Generation Conference*, 265–273. DOI: 10.18653/v1/W16-6644 (*In English*)

Obeidat, A. M., Ayyad, G. R., Sepora, T. and Mahadi, T. (2020). The tension between naturalness and accuracy in translating lexical collocations in literary text, *Journal of Social Sciences and Humanities*, 17 (8), 123–134. (*In English*)

Orešnik, J. (2002). Naturalness in English: some (morpho)syntactic examples, *Linguistica*, 42. DOI: 10.4312/linguistica.42.1.143-160 (*In English*)

Rachmawati, S., Sukyadi, D. and Samsudin, D. (2021). Lexical cohesion in the commercial advertisements of five Korean magazines, *Journal of Korean Applied Linguistics,* 1 (1), 29−44. (*In English*)

Rogers, M. (1998). Naturalness and Translation, *SYNAPS: Journal of Professional Communication*, 2 (99), 9–3. (*In English*)

Schramm, A. (1998). Tense and Aspect in Discourse, *Studies in Second Language Acquisition*, 20 (3), 433–434. https://doi.org/10.1017/s0272263198283069 (*In English*)

Schuff, H. & Vanderlyn, L. & Adel, H. & Vu, Th. (2023). How to do human evaluation: A brief introduction to user studies in NLP, *Natural Language Engineering*, 29, 1-24. DOI: 10.1017/S1351324922000535 (*In English*)

Serce, G. (2014). Relationship between naturalness and translations methods: Towards an objective characterization, *Synergies Chili*, 10, 139−153. (*In English*)

Siipi, H. (2008). Dimensions of Naturalness, *Ethics and the Environment*, 13 (1), 71−103. https://doi.org/10.2979/ETE.2008.13.1.71 (*In English*)

Sinclair, J. (1983). Naturalness in language, in Aarts, J. and Meys, W. (eds.), *Corpus Linguistics*, 203−210. (*In English*)

Talmy, L. (2000). *Toward a cognitive semantics*, vol. 2: Typology and process in concept structuring. Cambridge, Mass.: MIT Press (*In English*)

Thibault, P. J. (2011). First order languaging dynamics and second order language: The distributed language view, *Educational Psychology*, Vol.V, 32, 210–245. (*In English*)

Venuti, L. (1995). *The translator's invisibility*, Routledge, London and New York. (*In English*)

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter,D., Xia,F., Chi E., Le Qu., Zhou D. (2023). Chain-of-Thought Prompting Elicits Reasoning in Large Language Models, arXiv:2201.11903. https://doi.org/10.48550/arXiv.2201.11903 (*In English*)

Wilson, D. (1998). Discourse, coherence and relevance: A reply to Rachel Giora, *Journal of Pragmatics*, 29 (1), 57−74. (*In English*)

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q. and Artzi, Y. (2020). BERTscore: Evaluating text generation with BERT, *arXiv preprint arXiv: 1904.09675*. https://doi.org/10.48550/arXiv.1904.09675 (*In English*)

Zhou, J. and Bha, S. (2021). Paraphrase generation: A survey of the state of the art, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 5075–5086. (*In English*)

Zhu, Y., Lu, S., Zheng, L., Guo, J., Zhang, W., Wang, J. and Yu, Y. (2018). Texygen: A Benchmarking Platform for Text Generation Models, *arXiv preprint arXiv: 1802.01886*. https://doi.org/10.48550/arXiv.1802.01886 (*In English*)

Все авторы прочитали и одобрили окончательный вариант рукописи.

All authors have read and approved the final manuscript.

Конфликты интересов: у авторов нет конфликтов интересов для декларации.

Conflicts of interests: the authors have no conflicts of interest to declare.

Anastasia V. Kolmogorova, Doctor of Philology, Professor, Head of the Laboratory of Language Convergence, National Research University Higher School of Economics, St. Petersburg, Russia.

Анастасия Владимировна Колмогорова, доктор филологических наук, профессор, заведующий лабораторией языковой конвергенции НИУ ВШЭ Санкт-Петербург, Россия.

Anastasia V. Margolina, Machine Learning Engineer at Actum CBP d.o.o. Beograd, Serbia.

Анастасия Валерьевна Марголина, инженер машинного обучения, ДОО Actum, Белград, Сербия.