https://doi.org/10.48417/technolang.2025.04.13
Research article

Natural Language Processing and the Representation of Phenomenal Experience

Pavel Baryshnikov (✉) , Lolita Velis and Magomet Atakuev
Pyatigorsk State University, Kalinina Avenue, 9, 357532 Pyatigorsk, Russia
pnbaryshnikov@pgu.ru; lolitavelis@yandex.com; atakuevmagomet@gmail.com

Abstract

This article is an interdisciplinary study of phenomenal judgments through the lens of linguistic correlations using computational linguistics and data mining. The research focus includes perceptual judgments describing interoceptive and olfactory states, considered in the context of the theory of embodied cognition and Charles Sanders Peirce's theory of perceptual judgments. The authors demonstrate that the linguistic expression of interoceptive and olfactory experiences reflects a deep connection between the body, culture, and language, and also reveals culture-specific strategies for conceptualizing sensory experiences. Particular attention is paid to the comparison of "natural" and "synthetic" olfactory judgments that was generated using large language models (LLMs). The developed methodology allows for the identification of parametric differences in lexical diversity, syntactic complexity, and stylistic richness of olfactory descriptions. The conducted analysis confirms that olfactory experience has high semantic instability and polymorphism, which complicates its formalization and automated processing. Nevertheless, the use of modern NLP methods opens up new opportunities for the parameterization of phenomenal judgments and an in-depth study of their structural and cognitive features. The work is of interest from philosophical, humanitarian, and engineering points of view, offering a methodological toolkit for studying the properties of embodied consciousness using methods of computer processing of a natural language.

Keywords: Perceptual judgments; Embodied cognition; Interoception; Olfactory experience; NLP; Large language models; Phenomenal experience; Corpus analysis

Acknowledgment The research was funded by Russian Science Foundation No. 24-28-00540, https://rscf.ru/en/project/24-28-00540/

Citation: Baryshnikov, P., Velis, L. & Atakuev, M. (2025). Natural Language Processing and the Representation of Phenomenal Experience. *Technology and Language, 6*(4), 217-239. https://doi.org/10.48417/technolang.2025.04.13

УДК 004.8:81'27
https://doi.org/10.48417/technolang.2025.04.13
Научная статья

Компьютерная обработка естественного языка и представление феноменального опыта

Павел Барышников (✉), Лолита Велис и Магомет Атакуев
Пятигорский государственный университет, Пятигорск, проспект Калинина, 9, 357532, Россия
pnbaryshnikov@pgu.ru; lolitavelis@yandex.com; atakuevmagomet@gmail.com

Аннотация
Данная статья представляет собой междисциплинарное исследование феноменальных суждений через призму языковых корреляций с использованием методов компьютерной лингвистики и анализа данных. Исследовательский фокус включает в себя перцептивные суждения, описывающие интероцептивные (внутрителесные) и ольфакторные состояния, рассматриваемые в контексте теории воплощенного познания и теории перцептивных суждений Ч. С. Пирса. Авторы демонстрируют, что языковое выражение внутрителесного и ольфакторного опыта отражает глубокую связь между телом, культурой и языком, а также выявляют культурно-специфичные стратегии концептуализации сенсорных переживаний. Особое внимание уделено сравнению “естественных” и “синтетических” ольфакторных суждений, сгенерированных с помощью больших языковых моделей (LLM). Разработанная методология позволяет выявить параметрические различия в лексическом разнообразии, синтаксической сложности и стилистической насыщенности ольфакторных описаний. Проведённый анализ подтверждает, что ольфакторный опыт обладает высокой семантической нестабильностью и полиморфностью, что затрудняет его формализацию и автоматизированную обработку. Тем не менее, применение современных NLP-методов открывает новые возможности для параметризации феноменальных суждений и углублённого изучения их структурных и когнитивных особенностей. Работа представляет интерес как с философско-гуманитарной, так и с инженерной точек зрения, предлагая методологический инструментарий для изучения свойств воплощённого сознания методами компьютерной обработки естественного языка.

Ключевые слова: Перцептивные суждения; Воплощённое познание; Интероцепция; Ольфакторный опыт; NLP; Большие языковые модели; Феноменальный опыт; Корпусный анализ

Благодарность: Исследование было профинансировано Российским научным фондом № 24- 28-00540, https://rscf.ru/en/project/24-28-00540/

Для цитирования: Baryshnikov, P., Velis, L. Atakuev, M. Natural Language Processing and the Representation of Phenomenal Experience // Technology and Language. 2025. № 6(4). P. 217-239. https://doi.org/10.48417/technolang.2025.04.13

INTRODUCTION

The challenge of studying bodily experience through the lens of linguistic correlations using data science methods points to a new interdisciplinary area of research, in which common concepts and definitions are only just being formed. Natural language processing technologies (hereafter referred to as NLP) allow identifying new properties of the representation of bodily experience, both in regard to linguistic material and to philosophical epistemological. The proposed study results are obtained in the course of research at the intersection of several disciplines: analytical philosophy of consciousness, sensory linguistics, and computer data science.

The problematic dimension of this work lies in the paradigm of "embodied cognition" as opposed to classical computationalism (Baryshnikov, 2022). According to the body-focused approach, perception, action, and cognition are deeply interconnected and conceptualized during a person's physical interaction with the environment. In contrast, the computational paradigm argues that the neural structures of the brain perform computations similar to those performed by a computer, and that all mental states and cognitive processes can be reduced to basic symbols.

The results of empirical research in recent decades convincingly demonstrate the profound interconnection between human cognitive processes, bodily experience, and the characteristics of natural language both as a system and as an activity. Language plays a special role here, acting as a key instrument for reflecting this connection (Boroday, 2020). The unique property of language is that its use leaves an extensive "textual trace," which, given the current state of development of computer technology, can be recognized as a data set for computer analysis.

The proposed interdisciplinary approach expands the methodological arsenal and the existing understanding of the role of bodily experience in the formation of linguistic constructions. Data was collected while working with the Russian National Corpus (hereinafter referred to as the RNC). An own experimental collection of Russian literary texts was also created, and illustrative material from exotic languages, borrowed from studies in linguistic anthropology, was used (Kraska-Szlenk, 2023). Despite the new round of development of universalistic approaches (in connection with the revolutionary progress of the Large Language Models (LLMs), we believe that the relativistic approach allows gaining a deeper understanding of how bodily modes of sensation and processes of perception are reflected in the structures of language, and, conversely, how the structure of language influences the results of cognitive processes.

What can data science and natural language processing methods contribute to philosophical studies of phenomenal consciousness? This body-focused approach rejects traditional Cartesian dualism and sees cognition as something we do through our bodies as we interact with the environment. Some contemporary studies criticize the dichotomy often drawn between language and bodily experience, where language is seen as a mechanism structuring unstructured experiences. For example, Christoph Durt argues that such a view ignores how language and pre-linguistic behaviors are rooted in human interactions and how they shape our phenomenal experiences (Durt 2014). Today, the

analysis of language data allows for the extraction of useful information even in those practical areas where mental states are almost indescribable (Gamma & Metzinger, 2021).

Before the neural network revolution, there had been successful attempts to extract markers of mental states from text arrays using traditional algorithmic methods. The results of applying the NLP methods in behavioral science have been described in detail and systematized (Feuerriegel et al., 2025). The strength of these approaches is that computer analysis of texts allows for "visualizing" some objective correlations that are invisible when using other methods of analysis. Texts are data sets (Gentzkow et al., 2019), a kind of "digital traces" of the ways, in which linguistic signs are used. Statistical parameterization of texts allows extracting certain linguistic units and their contextual environment and identifying the features of word usage.

This approach opens up broad possibilities for comparative studies. Today, it is possible to extract linguistic data from phenomenal judgments without any serious technical difficulties, organizing them according to various characteristics: national language, era, age, profession, gender, literary authorship, situational context, emotionality, etc.

Particularly interesting is the comparison of phenomenal judgments that are natural (created by humans) or synthetic (created on retrained LLMs) (Muñoz-Ortiz et al., 2024). One of the stages of the proposed study includes precisely this kind of parametric comparative analysis based on the material of olfactory judgments.

WHAT DO PERCEPTUAL JUDGMENTS INDICATE?

According to Charles Sanders Peirce perceptual judgments are the first and immediate judgments about what a person perceives with their own sensory system at the moment (Peirce, 1978, 5.116). This type of judgment is the starting point or first premise for all subsequent acts of thinking and reflection.

There are several most significant criteria of perceptual judgments:

- Lack of control and impossibility of criticism. Peirce insists that the process of perceptual judgments formation is uncontrollable and, therefore, is not subject to criticism by logic since the subject is not able to consciously influence how they interpret the primary perception in the form of a judgment (Peirce, 1978, 5.55; 5.115).
- Distinction from percept. A perceptual judgment is not an absolute copy or direct reflection of a percept (perception of a visual image, sensation). It is a statement about the nature of the percept in propositional form.
- First premises of knowledge. Perceptual judgments are the fundamental type of judgments in relation to all other types of rational activity. All other judgments are theories whose validity is based on whether they are confirmed by perceptual judgments (Peirce, 1978, 5.116).
- Perceptual judgments as a case of abduction. Peirce considers perceptual judgments as a limiting case of abductive inferences (hypothetical conclusions). It is important to note that perceptual judgments are the result of a process of interpretation, not of passive reception of sensory signals. This process can be represented as an infinite series of abductions merging into a single act of perception (Peirce, 1978, 5.182; 5.184).

Perceptual illusions, where the interpretation of a figure changes (e.g., the Necker cube, the Penrose triangle), indirectly indicate the kinship of perception and abduction. Such a representation of the perceptual process brings it closer to the modern cognitive theory of "predictive processing" (Mudrik et al., 2025).

It is also important to note that "perceptual judgments" correspond to iconic signs, which are linked to each other by a similarity relation (Peirce, 1978, 5.119). In other words, iconic signs directly stand for the qualities or features described in perceptual judgments.

As a result, based on the Peirce's definition, we can say that a "perceptual judgment" is a statement in propositional form about the nature of a percept, directly given in experience. Perceptual judgments are closely related to the process of perception, which does not imply passive processing of sensory signals of one of the modalities. This type of judgment is also based on the interpretation of the percept itself for the representation and an understanding of the surrounding reality. Thus, such expressions as „my stomach is all balled up", "there's a squeezing feeling in my chest", "my chest feels tight", "citrus scent", "smells like freshly cut grass" are examples of perceptual judgments, since they contain a subjective interpretation of the perceptual signal.

In the context of body-oriented approaches, metaphor is seen as a conceptual and structuring aspect of cognition. Based on the Theory of Conceptual Metaphor (TCM) (Casasanto, 2017; Gibbs, 2004; Zhao, 2023), proponents of "embodied cognition" construct arguments in support of the view that the linguistic representation of bodily experience is extracted from the interaction of the embodied agent with the environment. An earlier version of TCM formed a universalistic view of bodily experience, while representatives of a modern iteration of this theory emphasize the culture specific features of metaphor (Yu, 2020). This approach allows asserting that the properties of linguistic conceptualization are not simply the result of abstract representations, but are extracted from sensorimotor activity and the interaction within the "body – language – culture" triad.

Interoception and olfactory experience

Interoception is a critical aspect of perceiving and processing signals emanating from the intra-body space (e.g., hunger, thirst, fever, pain, etc.). It is also an important mechanism for maintaining the internal physiological state of the body. In recent years, interoception has become an important research object in psychology, neuroscience, neurophysiology, and clinical medicine (Murphy, 2024). Modern empirical data in the field of cognitive science indicate that interoception plays a key role in social cognition, emotional experience (Feldman et al., 2024), self-awareness (Seth & Tsakiris, 2018), and other high-level mental processes.

In this study, we consider the influence of linguistic mechanisms for conceptualizing intra-body experience, since there is an ontological gap between the objective bodily event and the experience of interoception in the first-person perspective. This difference points to the importance of considering not only the neurophysiological correlates of interoception, but also the culturally conditioned forms of conceptualization, by which the speech agent formulates judgments about their own bodily experience. The

interoceptive and olfactory narrative reflects cultural features that influence the qualitative characteristics of subjective experience.

It is generally accepted that the conceptual apparatus of olfactory experience is referential, i.e., referring to an object that emits a certain smell. For example, “smells like a banana,” “the smell of a rose,” “smells like asphalt after rain.” The specific linguistic representation of olfactory experience presupposes the existence of an external object, to which the concept refers. In this regard, it is difficult to talk about the existence of an adequate language, in which independent olfactory concepts exist, as, for example, in the language of visual experience (“red”, “white”, “purple”, etc.).

In the context of cross-cultural linguistic research (Zhang, 2011; Shaules, 2020), telling arguments have been put forward against the thesis that linguistic constructions expressing olfactory experience are poorly suited for naming olfactory phenomena. To prove this, we can give some examples from the Jahai language. This is a Mon-Khmer language spoken by nomadic hunter-gatherers on the Malay Peninsula and it has a rich olfactory vocabulary. This linguistic system includes words for odors that are not associated with the source of the odor (Majid & Burenhult, 2014). For example, in Jahai one can find concepts expressing

- “the smell of blood that attracts tigers (e.g., crushed lice, squirrel blood)”
- “to be fragrant (e.g., various types of flowers, perfume, soap)”
- “the smell of mustiness (e.g., old shelter, mushrooms, stale food)” (Ibid., p. 269).

Unlike English speakers, who tend to use descriptive constructions that point to the actual source of the smell, Jahai speakers tend to use abstract concepts to describe olfactory and visual experiences. These examples demonstrate how cultural practice and linguistic structure influence cognitive perception.

Among the Jahai people, the body serves as a kind of metaphorical template for describing the physical world. For instance, local neighborhoods, houses, and landscapes are described through bodily concepts (Burenhult, 2006). It is noteworthy that this language is distinguished by its referential discreteness in relation to somatic concepts. A consequence of this is the fact that this language does not have a term for “face” or “mouth,” but has a wide range of categories for the constituent parts of the face: frontal tubercle, upper/lower lip, baby tooth, root of the nose (refers to the wrinkles between the eyebrows), etc. (Burenhult, 2006, p. 167).

The above example suggests that perceptual judgments, embodied cognition, and culturally specific conceptualizations of bodily and sensory experience reveal the mechanisms, through which language serves as a tool not only for describing but also for shaping phenomenal experience mediated by the body and culture. The question arises: Can one with current methods teach Large Language Models (LLMs) the principles required for understanding this deep and implicit relationship? Will the generated artificial perceptual judgments have the metaphorical features of bodily conceptualization?

STATISTICAL PARAMETERIZATION OF OLFACTORY JUDGMENTS

Today, cognitive aspects of olfactory experience attract representatives of various disciplines – from the philosophy of consciousness and cognitive linguistics to applied psychology and AI methodology (Hörberg et al., 2022; Jraissati & Deroy, 2021; Martina, 2023; Young, 2016). Natural language processing (NLP) technologies can help identify the linguistic properties of bodily experience and examine their connections to central epistemological categories in philosophy.

Modern computational linguistics methods allow conducting a wide variety of textual studies. Here are a few possible approaches to the study of olfactory vocabulary:

1. Text analysis by search word (SW). Here, morphosyntactic analysis is used to collect dependent words, i.e. the contextual environment associated with the search word and a related dependent word, e.g., "smell" of „burning." This method is convenient when it is necessary to view the frequency of the search word or to derive the frequency of the dependent words.
2. Text analysis for the presence of olfactory code using NER (Named Entity Recognition) or search by a named entity. This method requires a pre-trained model for recognizing named entities. This method is convenient when it is necessary to recognize all possible olfactory descriptions in the text at once. However, there are several significant drawbacks – firstly, the very process of model training, which requires a large amount of labeled data, and secondly, the complexity of verifying the results obtained.

The proposed practical results of the study were obtained through research at the intersection of several disciplines: analytical philosophy of consciousness, sensory linguistics, and computer data science. The study used NLP technologies for the analysis of the text corpus (frequency analysis, clustering, topic modeling). Such an interdisciplinary approach expands the philosophical methodological arsenal and the existing understanding of the role of bodily experience in the formation of linguistic constructions. Data were collected from the Russian National Corpus (RNC), namely our own experimental collection (hereafter referred to as LC_69) of 69 prose texts of Russian literature (19th–20th centuries).[1] It consists of 690,000 sentences and includes more than 3,000 descriptions (direct and indirect) of olfactory experience.

The relational database consists of five tables linked by an external key according to the M2M principle ("Many-to-Many" is a database relationship where multiple records in one table link to multiple records in another, typically resolved using a junction table.).:

[1] The collection of texts is presented here: https://github.com/kagort/tolstoy-words-local/blob/PN_test/verbal_forms/data/processed/dictexts.csv

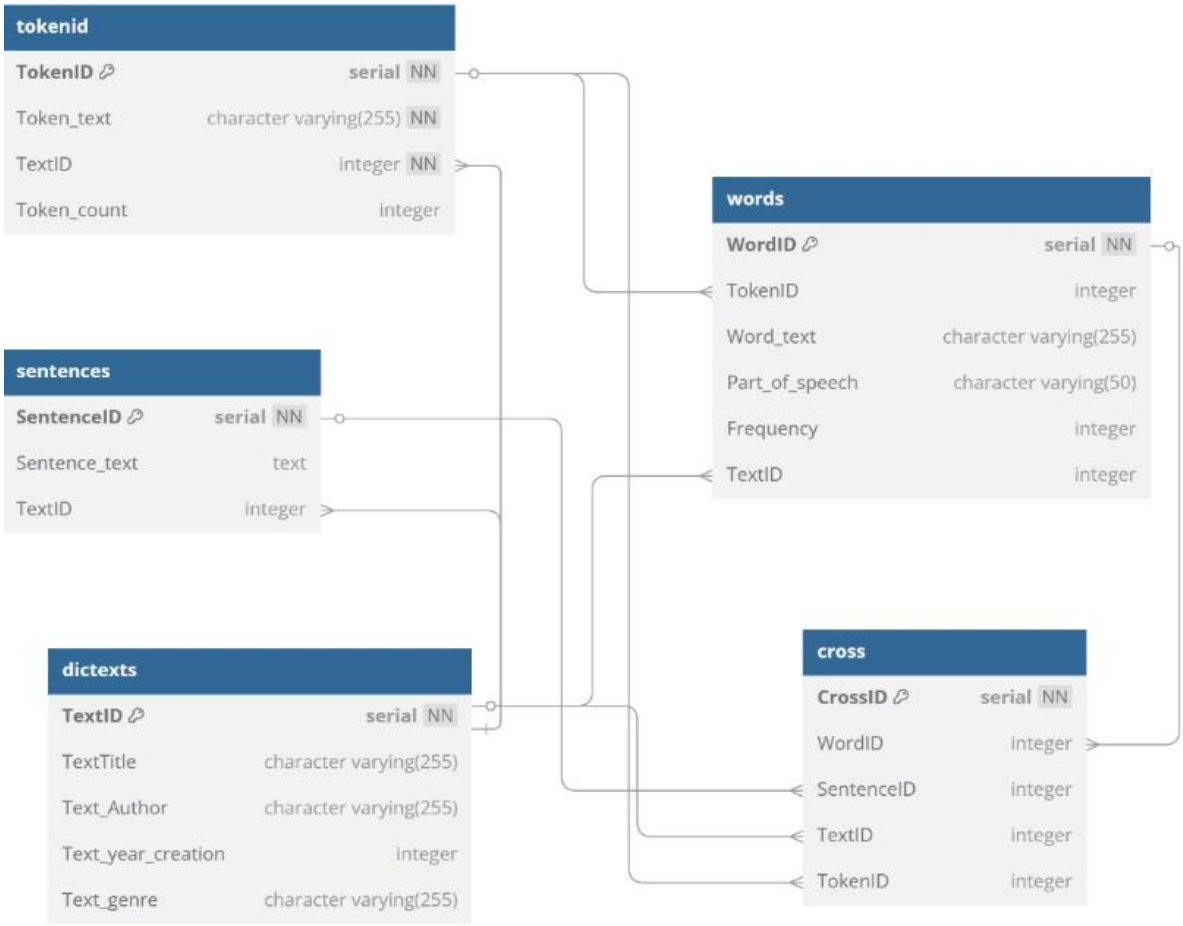
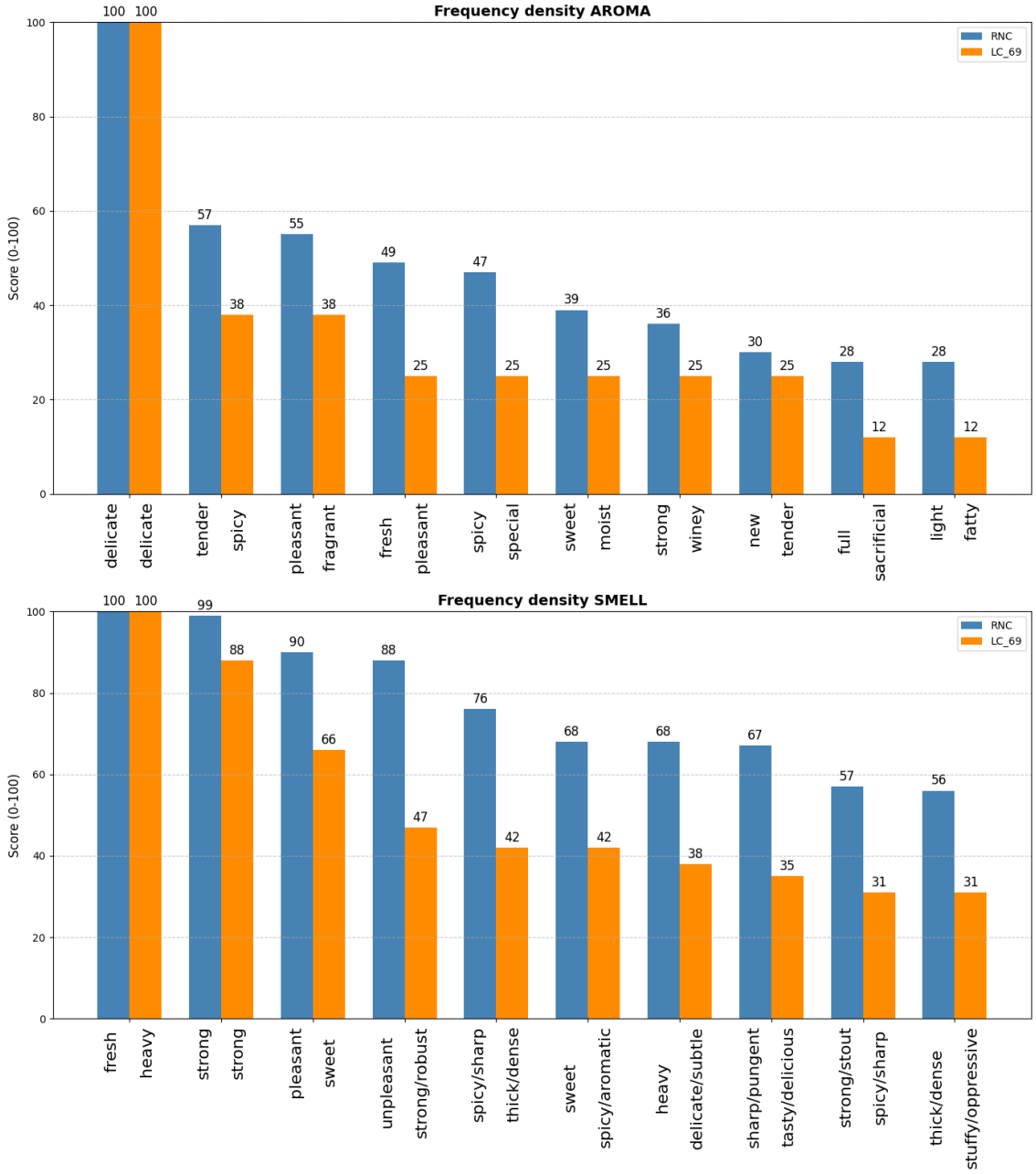


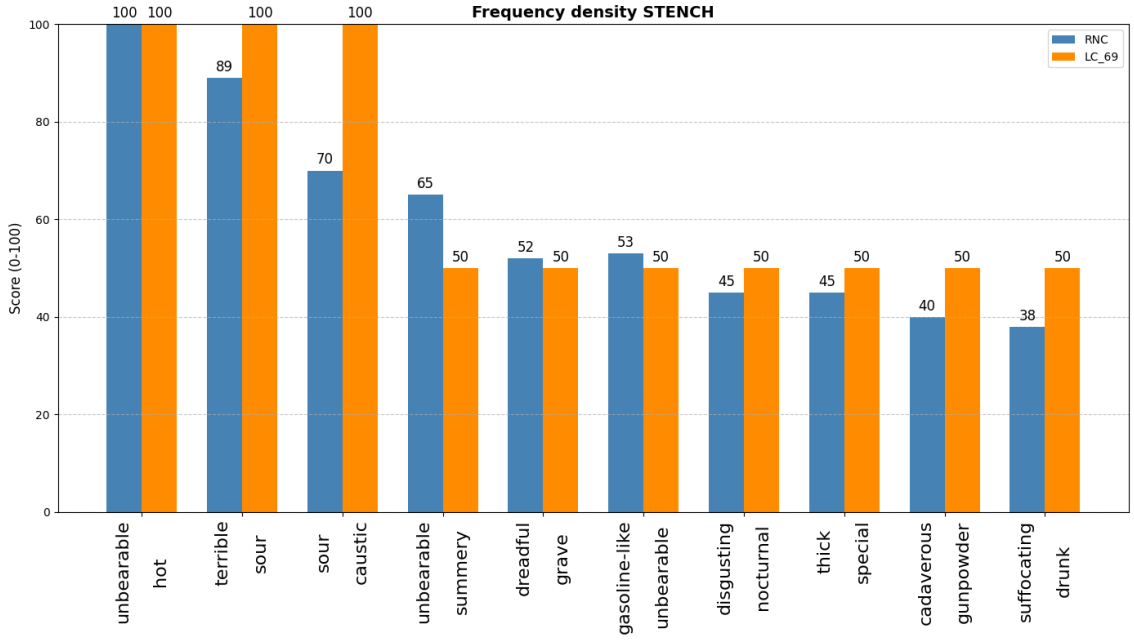
Fig. 1. ERD structure of the database of texts from Russian literature

The final list of the top 5 search words includes: “smell” (запах [zapakh]) (1019), “aroma” (аромат [aromat]) (92), “spirit” (дух [dukh]) (74), “stench” (вонь [von]) (57), “stink” (смрад [smrad]) (39). A problem arose with the inconsistent machine lemmatization of the word “perfume” (духи [dukhi]): only by way of manual verification it was possible to determine how many times out of 74 occurrences in the text the word “perfume” (духи [dukhI]) was used and how many times the word “spirits” (духи [dUkhi]) (they are spelled identically but differ in pronunciation and meaning). Taking this into account, the final sample was limited to 4 search words: “smell,” “aroma,” “stench,” and “stink.” For each of the words, its contextual environment was selected according to the following parameters: N+ADJ, N+N.gen (common noun phrase structures: noun + attributive adjective, and noun + noun in the genitive case).

The idea of comparing absolute values of the frequency of the most representative collocates in the RNC and in LC_69 is dictated by the need to check the dependence of the curve on the frequency density. In other words, the larger the corpus, the smoother should be the drop in values from maximum to minimum.

Table 1. Noun + adjective (RNC – blue, LC_69 – yellow)





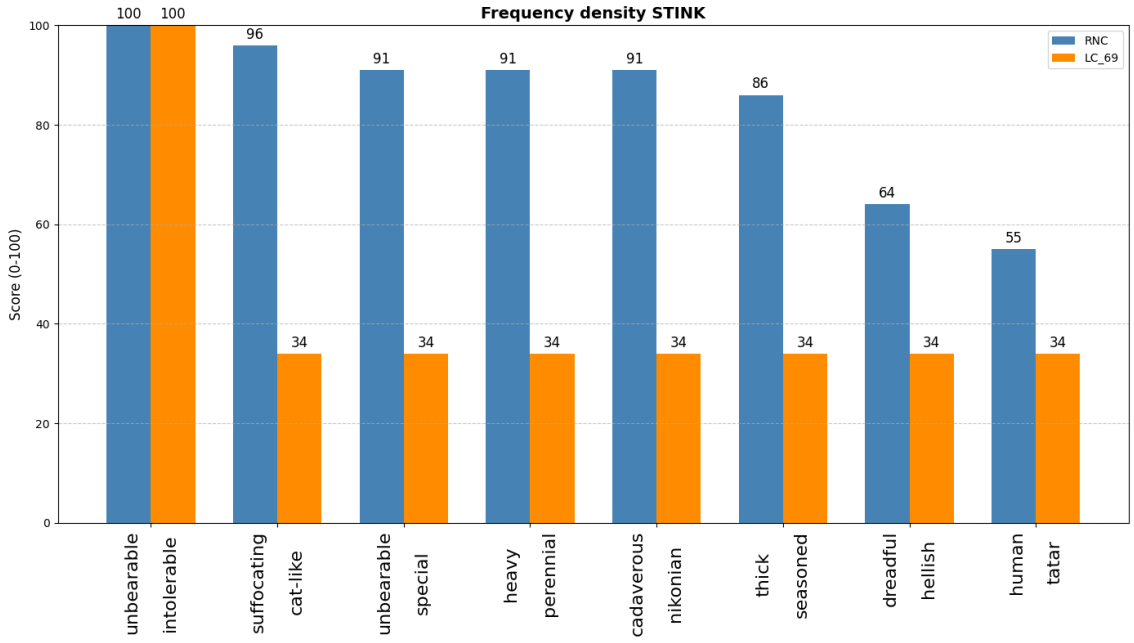


Table 1 shows that with a larger volume, the curve of the collocates (common word pairs) decline is smoother. The only exceptions are cases with the “aroma” lexeme, which may indicate the specificity and contextual limitations of the use of the lexeme itself. Moreover, in the sample of literary texts, the decline is even more pronounced, which may also indicate that “aroma” is not very popular among early classics.

More representative results are shown by the method of lexical series clustering both for the entire corpus and for individual works. As part of the experiment, an algorithm for semantic clustering of lexical units was implemented based on pre-trained

language models and machine learning methods. A table with 662 collocates of the keywords “smell” and “aroma” and a data set of the following structure was submitted for analysis:

Table 2. Collocates (N+ADJ, N+N.gen.) and their rates of frequency

Russian Collocation	English translation	Frequency
тяжёлый запах	heavy smell	91
запах духов	smell of perfume	78
сильный запах	strong smell	76
сладкий запах	sweet smell	58
крепкий запах	intense smell	55
запах пота	smell of sweat	53
запах цветов	smell of flowers	49
тонкий запах	subtle scent	44
густой запах	thick smell	40
пряный запах	spicy smell	35
горький запах	bitter smell	32
запах крови	smell of blood	32
запах травы	smell of grass	31
…	…	…

At the first stage, the words were prepared for analysis: for more accurate processing, the data were expanded taking into account the frequency rate of each phrase. After that, each phrase was matched with a numerical representation (a vector) reflecting its meaning in the linguistic context. For this, the neural network model “intfloat/multilingual-e5-large” was used, capable of capturing subtle differences in the meaning of expressions by constructing embeddings for each expression.

At the next stage, the phrases were combined into groups (clusters) based on thematic proximity. One of the popular machine learning methods, the KMeans algorithm, was used as a basis. To determine how well the resulting groups, correspond

to the natural structure of the data, a special silhouette coefficient was used. This is a kind of "quality assessment" of the grouping, which helps to understand how organically the data is divided into clusters. Each group received a name reflecting its semantic content, based on the most characteristic words.

The clustering results were visualized in two-dimensional space using the t-SNE algorithm. This allowed us to see for different phrases how close or far from each other they were in meaning. Cluster distribution diagrams were also generated. However, it should be noted that the results are not always reproducible precisely. When the algorithm is re-run, the cluster structure may change slightly, although the data and parameters remain the same. This is the effect of the operating principles of the model.

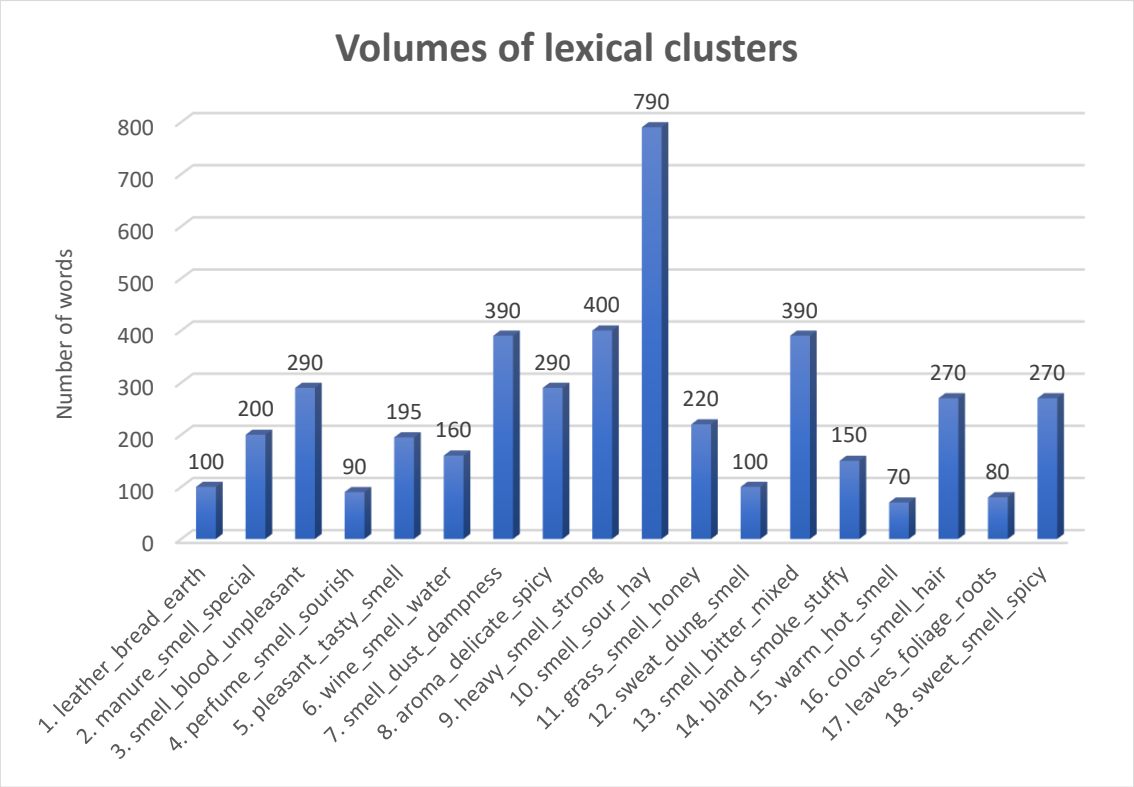


Fig. 2. Volumes of lexical clusters

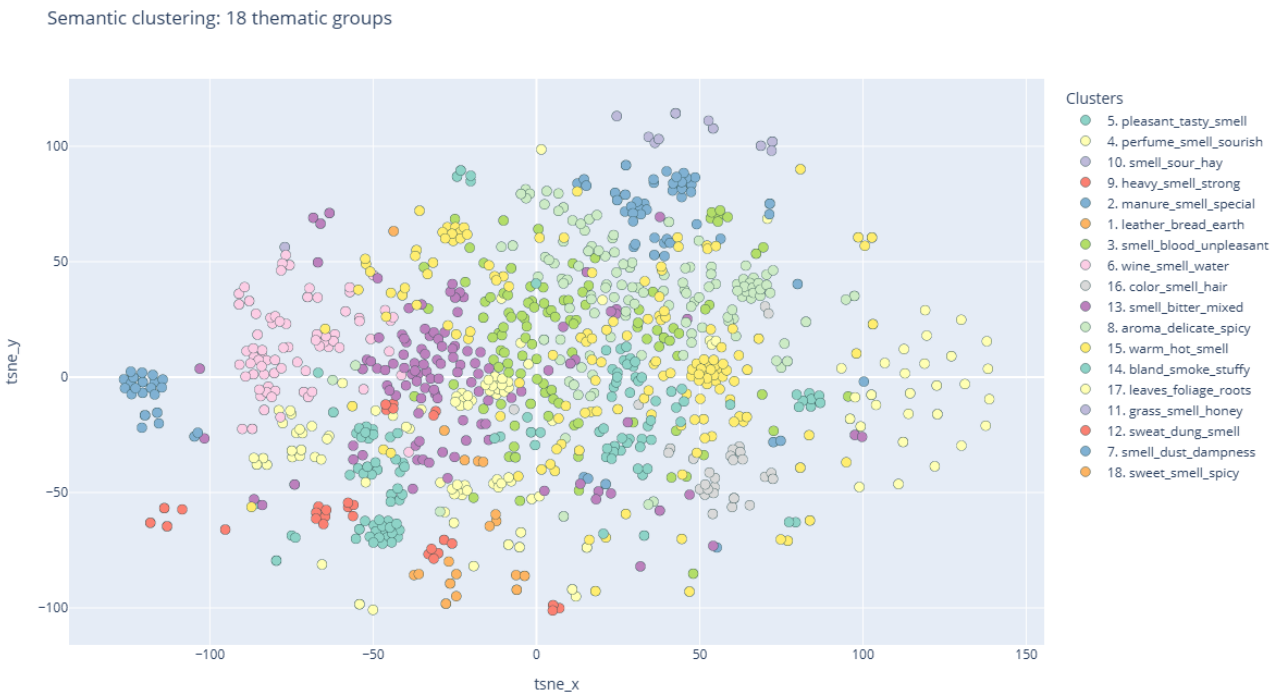


Fig. 3. Volumes of lexical clusters

This method allows us to identify a unique olfactory trace throughout the text corpus. Cluster 9 named “smell sour_hay” (the slightly sour smell of aged hay) obviously dominates. “Sour” and “hay” are included in the cluster name due to the largest number of occurrences of one of the collocates. To understand the conditional logic of the model, we decided that it was necessary to build an associative map for a given lexical series related to the concept of smell. We extracted data from the N.gen cluster and, based on these formed strings of pre-calculated wordembeddingss. Then these strings were transformed back into NumPy arrays (NumPy arrays — numerical arrays for fast computations in Python). After that, a matrix of cosine similarities between the embeddings of the selected words was calculated, based on which pairs of words with a semantic proximity level above a given threshold (0.7) are formed. Next, an undirected graph was built using the NetworkX library, where nodes represent words, and edges represent significant associative links between them.

It is also important to note that if there is a dominant cluster with volume significantly larger than the average, then its content will appear the least associative. A dominant cluster becomes a "catch-all" category – so broad that the things inside it don’t strongly remind us of each other. Hence, its content appears less tightly connected by meaning, perception, or usage.

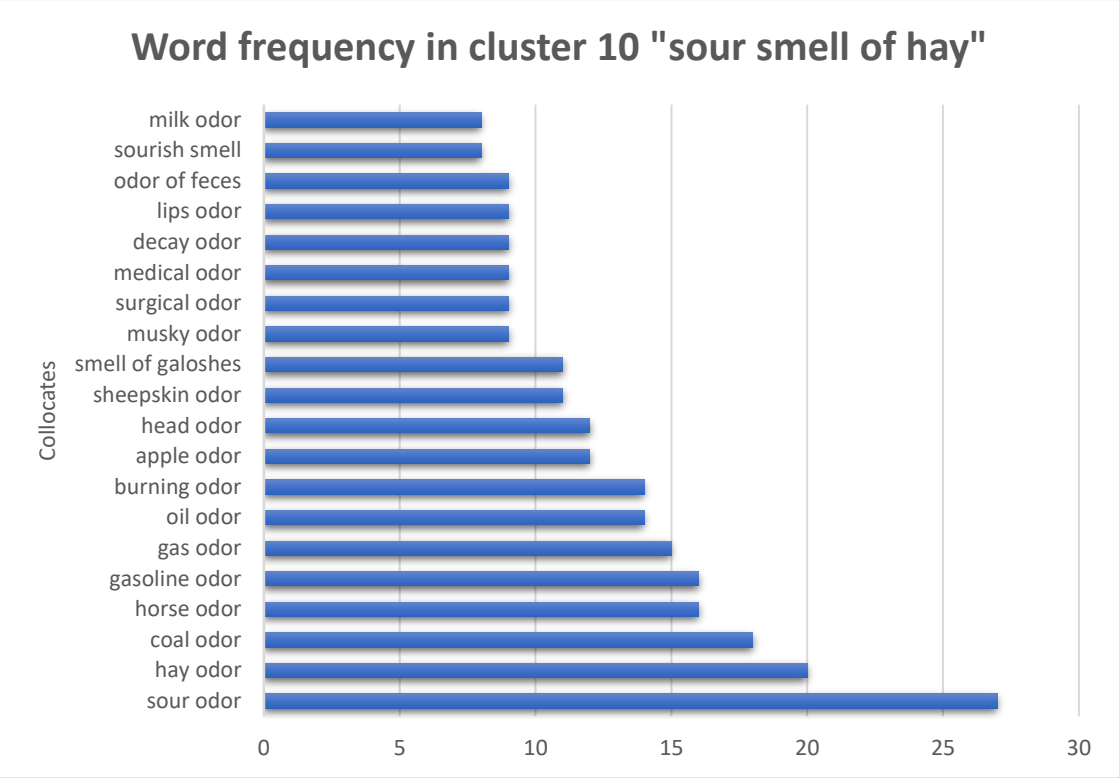


Fig. 4. Word frequency in cluster 10

Automatic thematic grouping of words included in the cluster so far leads to unsatisfactory results. If we group the collocates "hay smell", "sour smell", etc., then the graph density becomes very close to 1, i.e. the model recognizes all data as belonging to the "smell/aroma" group. If we analyze a list of dependent words, then the graph density disintegrates.

If we analyze the content of cluster 9, we obtain the following set of 218 dependent lexemes, which can be thematically grouped as follows. See the below Table.

Table 3. Results of thematic grouping of lexemes from Cluster 9 using the expert assessment method

Category	Category composition
1. Smells / Descriptions of odors	*угля (coal), кислота (acid), кал (feces), кошачий (cat's), гошпитальным (hospital (adj., obsoletism)), эфира (ether), дегтярный (tar (adj.)), ментол (menthol), чесночный (garlic (adj.)), перегар (booze breath), кизячного (dung (adj.)), клоака (cesspool), гумен (manure pits), извержение (eruption), подмышка (armpit), моча (urine), йодоформ (iodoform), хирургический (surgical), больничный (hospital (adj.)), гари (soot), ржи (rye), хрена (horseradish), калошный (galosh), тлена (decay), асфальт*

	(asphalt), угарный (carbon monoxide), звериный (animal (adj.)), ил (silt), сенной (hay (adj.)), гвардейский (guards (adj.)), мужицкий (peasant (adj.)), машины (machine's), уксус (vinegar), хмельной (intoxicating), мазь (ointment), мех (fur), миндальный (almond (adj.)), мускус (musk), клоп (bedbug)
2. Natural/plant sources of odors	*герань (geranium), донник (sweet clover), туберoза (tuberose), можжевельник (juniper), васильковый (cornflower (adj.)), чабрец (thyme), чернобыль (полынь) (wormwood), кора (bark), березняк (birch), луговой (meadow (adj.)), глина (clay), зола (ash), воск (wax), чеснок (garlic), лук (onion), капуста (cabbage), картофель (potato), клубника (strawberry), яблоко (apple), груша (pear), фрукт (fruit), огурец (cucumber), сено (hay), травяной (herbal), фиалковый (violet (adj.)), багульник (wild rosemary)*
3. Food / Drinks / Dishes / Cuisine	*яйцо (egg), сок (juice), котлета (meat ball), кушанье (meal), блюдо (dish), кабак (pub (obsoletism)), маслина (olive), квасной (kvass (adj.)), бражный (molasses mash-like), браги (molasses mash), столовый (table (adj.)), кашица (gruel), каша (porridge), каравай (round bread (obsoletism)), кухня (kitchen), кухонный (belonging to kitchen), еда (food), корм (food for animals), ветчина (ham), селёдка (herring), масло (butter), жир (fat), молоко (milk)*
4. Objects / Places / Phenomena	*утюг (iron), юбка (skirt), бельё (linen), занавеска (curtain), шлак (slag), упряжь (horse harness), почка (bud), склянка (bottle (obsoletism)), картуз (cap (obsoletism)), рот (mouth), губы (lips), головы (heads), испражнение (excrement), кот (tomcat), кулиса (drop cloth), воз (cart (obsoletism)), машина (car), дорога (road), пыль (dust), пепел (ashes), лазарет (infirmary (obsoletism)), аптека (pharmacy), больница (hospital), церковь (church), рабфак (workers' faculty (obsoletism)), склеп (crypt), ус (mustache), баз (base), железо (metal), железный (metal (adj.)), железнодорожный (railway (adj.))*
5. Chemistry / Medicines / Reagents	*кислота (acid), эфир (ether), йод (iodine), йодоформ (iodoform), нафталин (naphthalene), хлор (chlorine), керосин (kerosene), бензин (gasoline), угарный (carbon monoxide), уксус (vinegar), спирт (alcohol), одеколон (cologne), клейстер (flour paste), клей (glue), ржавчина (rust), смола (resin), сажа (soot), одеколонный (cologne (adj.))*
6. Animals / Animal Smells	*коровий (cow's), кобыла (mare), кот (tomcat), котлета (meat ball), конь (horse (male)), лошадь (horse (female), кошка (cat (female)), тюлень (seal), мышь (mouse), крыса (rat), муравей (ant), паук (spider), комар (mosquito), клоп (bed bug), пчела (bee), оса (wasp), муха (fly), собака (dog), змея (snake), свиной (pig (adj.)), конский (horse (adj.)), лошадиный (equine (adj.))*

7. Colors / Textures / Properties	*белёный (bleached), зелёный (green), кисленький (slightly sour), кисловатый (sourish), кислый (sour), бархатистый (velvety), лакированный (varnished), томительный (languid), густой (thick), жидкий (liquid), сладкий (sweet), горький (bitter), острый (spicy)*
8. Metaphors / Abstractions / Images	*юность (youth), время (time), воспоминание (memory), страх (fear), радость (joy), любовь (love), дом (home), детство (childhood), старость (old age), весн (spring), лето (summer), осень (autumn), зима (winter), вечер (evening), рассвет (dawn), ночь (night)*

The results of experiments with olfactory vocabulary demonstrate limited functionality, which indicates, on the one hand, the technical limitations of the NLP models used, and on the other hand, the peculiarities of olfactory experience as one of the types of sensory perception. Olfactory experience is characterized by high individual variability and emotional coloring. The perception of smells is closely connected to associative memory, which makes the linguistic expression of the olfactory profile of perception extremely unstable and polymorphic – a smell can be described through objects, metaphors, textures, colors, as well as abstract images and affective states.

Unlike visual or auditory perception, which have stable categorical systems (colors, shapes, volume, timbre), olfaction does not have a developed conceptualized structure, which complicates its verbalization and formalization. The language of smells turns out to be extremely conditional and therefore is difficult to classify by machine learning tools focused on the rational-conceptual organization of language. The presented results can be analyzed in two projections:

Negative projection. The poor quality of clustering points to a fundamental characteristic of olfactory vocabulary, namely the dependence of this perceptual mode on bodily experience. This characteristic is an important factor limiting the possibilities of automatic analysis and requiring the development of specialized approaches that take into account the somatic and affective nature of odor perception.

Positive projection. The clustering and thematic modeling of olfactory judgments requires a different engineering approach. For example, labeled training datasets of olfactory judgments (see Table 3) will help to further train LLMs in order to navigate the conceptual blurriness of olfactory judgments based on statistical relationships and patterns.

After the first stage of additional training, sets of “synthetic” olfactory judgments were obtained, for the analysis of which additional metrics were developed. The use of these metrics should be aimed at obtaining a verifiable result that allows distinguishing the features of “natural” and “machine” types of conceptualizations.

OBTAINING "SYNTHETIC" OLFACTORY JUDGMENTS

One of the stages of the experiment was further training of the model for the task of generating perceptual judgments with olfactory description.

This process was carried out on the Qwen 2.5-7B model, namely, on its adapted version for the Russian language (Hugging Face: RefalMachine/ruadapt_qwen2.5_7B_ext_u48_instruct). The initial data were sentences containing olfactory vocabulary, taken from Russian classical literature. The model was further trained using the Unsloth library, which supports dynamic 4-bit quantization (Unsloth - Dynamic 4-bit Quantization).

Along with the quantization of the model, LoRA (Low-Rank Adaptation) was used. LLMs (Large Language Models) are pre-trained on a large corpus of general data and can then be customized for specific task(s). Due to the large size of the LLM, full customization of all parameters becomes prohibitively expensive. Using LoRA reduces the number of parameters to be trained, which results in reduced training time and GPU memory usage while maintaining the quality of the output data. Thus, additional training of the Qwen 2.5-7B model using quantization and LoRA allowed us to adapt a large pre-trained model to a specific task and dataset without retraining the entire model. LoRA is embedded only in certain layers of the model, which reduces the number of parameters to be trained, reduces GPU memory requirements, speeds up training, and improves the accuracy of the model for specialized tasks.

Each sentence from the corpus was designed as an instruction + response, using the Alpaca style prompt structure. This dataset organization allows the model to correctly extract the necessary information from the instruction. The texts were then converted into a list of dictionaries and loaded into the HuggingFace Dataset format.

For additional training, SFTTrainer (supervised fine-tuning) from the Unsloth library was used, which made it possible to efficiently use the available video memory (15 GB) and significantly reduce computational costs while maintaining the efficiency of training.

After completing the additional training stage, several inference runs were conducted to evaluate the generative capabilities of the model. The values of the temperature parameter (which regulates the randomness of model output) were iterated from 0.40 to 0.85 with a step of 0.05.

For each temperature value, the model generated 100 sentences using a random instruction:

> *"You are an author writing about your olfactory experience. Your task is to create descriptions of smells, immersing the reader in your sensations. Use olfactory markers of your choice for description: aroma, smell, stench, fragrance, fetor, odor, stink, incense, etc."*
> *"Write a literary text in which smells reveal the inner state of the character. Use metaphors, comparisons, and unexpected images. Don't limit yourself to the word 'smell', use words like 'aroma', 'stench', 'fragrance', 'fetor', 'odor', 'stink', 'incense-'."*

> *"Convey the atmosphere of the scene through smells: let them reflect the characters' feelings, evoke associations, or heighten tension. Write figuratively and metaphorically."*

The generated texts were saved in separate files for each temperature value. According to the results of manual analysis of the texts, the most meaningful were the sentences composed at a temperature of 0.4, which formed the basis of the dataset of generated texts with olfactory markers.

As a result, a dataset of 4323 synthetic or machine descriptions with contexts was obtained. In fact, the expert assessment method allowed us to identify repetitions in the structure, poor lexical compatibility, and stylistic errors.

For example:

(1) В дальнем углу комнаты, где свет почти не достигал, можно было уловить слабый запашок плесени. Он напоминал о том, что время не щадит ничего, даже самые изысканные ароматы могут быть омрачены неприятными запахами.
(In the far corner of the room, where the light barely reached, a faint smell of mold could be detected. It was a reminder that time spares nothing, even the most exquisite aromas can be overshadowed by unpleasant ones.)

(2) В воздухе витал аромат цветущих яблонь, смешиваясь с медовым запахом полевых цветов.
(The air was filled with the scent of apple blossoms, mingling with the honeyed scent of wild flowers.)

(3) В воздухе витает аромат свежескошенной травы, смешанный с благоуханием цветущих садов.
(The air is filled with the scent of freshly cut grass, mixed with the fragrance of blooming gardens.)

(4) Но не всё было так сладко. Вдали, на окраине деревни, можно было уловить зловоние гниющих отходов, которое резко контрастило с общей картиной утренней свежести. Этот смрад, казалось, проникал в каждую щель, напоминая о несовершенстве мира.
(But not everything was so sweet. In the distance, on the outskirts of the village, one could detect the stench of rotting garbage, which contrasted sharply with the general picture of morning freshness. This stench seemed to penetrate into every crack, a reminder of the imperfections of the world.)

Our tasks included developing statistical evaluation metrics that could identify regular parametric differences between "natural" and "synthetic" olfactory descriptions. An algorithm was created that implemented a complex analysis of lemmatized tokens, capable of extracting frequency characteristics from text data using psql and pandas. The program builds regular expressions to search for all forms of a word, calculates the frequency of tokens, the number of sentences in which they occur, and analyzes dependent parts of speech and sentence lengths, subsequently saving the results in tables.

Here we will present only the final results:

Table 4. Comparison of parameters of “natural” and “synthetic” olfactory descriptions

Indicator	Natural Tokens	Synthetic Tokens
Unique tokens	34	19
Average number of token occurrences	706.5	516.2
POS variety (on average per token)	9.68	8.95
VERB_HEAD	2809	1025
NOUN	2106	701
ADJ	1454	295
ADP	546	93
VERB	500	431
VERB_HEAD	2809	1025

The corpus with “natural” tokens has higher lexical diversity, average word length, and variety of grammatical roles. This may indicate greater stylistic or semantic richness. In all POS categories, the corpus of “natural” tokens demonstrates significantly higher absolute values. The difference is especially noticeable in adjectives (ADJ), nouns (NOUN), and adpositions (ADP), which indicates greater descriptiveness and syntactic ramification. It is worth mentioning the problem of back translation which arises from the fact that a decrease in lexical variability may be associated with the features of the model architecture. Most modern multilingual LLMs are trained on English-language data. Therefore, when working with other languages, a “back translation effect” occurs frequently, namely when the model first generates a response in English and then translates it into the query language (Zhang et al., 2020). As is known, English has one of the richest vocabularies among the languages of the world, while Russian is more complex in morphosyntactic terms. This may explain the great variability in lexical compatibility and the general lexical richness of the Russian language, especially in limited thematic domains (Mariko, 2025).

The most obvious general characteristic of LLMs is manifested in the peculiarities of the average sentence length. Our hypothesis was that the length of sentences with the selected tokens differs significantly in the generated and natural texts. To test the hypothesis, we used the classical Walsh t-test (Shaules, 2020), since it can be used on small samples to compare average values with different variances. For 10 of the 12 studied pairs of tokens, the differences in the average sentence length are statistically significant ($p < 0.05$). The average sentence length in synthetic texts was 14–17 words

versus 21–37 words in natural texts. The variance of sentence length in synthetic texts was significantly lower, indicating a smaller variety of syntactic constructions.

It is evident that the language model produces sentences that are shorter and also syntactically more uniform than natural human texts. This reflects a fundamental feature of LLMs: they are optimized to reduce prediction uncertainty and enhance local coherence, which leads them to favour common, predictable syntactic structures. In contrast, human descriptions of embodied or sensory experiences often exhibit irregular, fragmented, or metaphor-rich syntax, reflecting the difficulty of putting pre-reflective, non-propositional states into words. The model, however, replaces this expressive variability with repetitive templates (e.g., "It smells like X" or "It has a X smell"), thereby smoothing out the linguistic traces of phenomenological effort. As a result, while the generated text appears fluent and coherent, it lacks the cognitive and affective complexity that characterizes authentic perceptual reports. We can conclude that, this syntactic simplification is therefore not merely a stylistic limitation, but a systematic bias that masks the richness and ambiguity of embodied experience behind a surface of grammatical regularity.

CONCLUSION

Language judgments related to the perception of odors and the experience of intra-corporeal (interoceptive) states are an important research object. Semantic properties expressed in statistical parameters allow revealing the features of bodily conceptualization through the "prism" of a particular linguistic picture of the world. The problem of sensory experience verbalization is well covered in modern research (Nagornaya, 2014; Winter, 2019). Nevertheless, natural language processing (NLP) technologies often ignore bodily aspects, which creates a certain gap in the adequate interpretation of sensory metaphors and socio-cultural contexts. In this study we developed and applied analytical methods to odor-experience text collections that enable its systematic detection – revealing measurable parameters of lexical, syntactic, and stylistic divergence. Perhaps, further developments will help to reduce the detected parametric gap. We see such an applied interdisciplinary approach as promising for both philosophical generalizations and engineering knowledge.

In defending the body-oriented approach in discussions of the phenomenal content of consciousness, we have substantiated that olfactory experience in natural language is described through a wide range of lexical collocations, including objects, emotions, colors, and abstract images, which indicates the polymorphism of experience and its emotional richness. Judgments describing phenomenal experience represent access to a complex system of relationships between individual introspective reports and the socio-cultural component of language habits. These relationships and their context-dependencies remain inaccessible (or available only in a distorted form) to machine learning methods. Modern computer methods of natural language processing can help us implement the principles of strict parameterization of phenomenal judgments. This allows performing a contrastive analysis – of vocabulary, grammar, stylistics, and other language levels – of natural and synthetic perceptual descriptions, those found in the corpus and

those generated by a LLM after additional training. A comparative analysis of human- and machine-generated olfactory descriptions reveals tendencies by LLMs of statistical generalization of lexical composition along with syntactic simplification – strengthens the position of those who endorse philosophical anti-computationalism.

REFERENCES

Baryshnikov P.N. (2022). *Computational Models of the Mind: From Code to Meaning*. URSS.

Borodai, S. Yu. (2020). *Language and Cognition: Introduction to Postrelativism*. YASK Publishing House.

Burenhult, N. (2006). Body Part Terms in Jahai. *Language Sciences, 28*(2–3), 162–180. https://doi.org/10.1016/j.langsci.2005.11.002

Casasanto, D. (2017). Relationships between Language and Cognition. In B B. Dancygier (Ed.), *The Cambridge Handbook of Cognitive Linguistics* (pp. 19–37). Cambridge University Press. https://doi.org/10.1017/9781316339732.003

Durt, C. (2014). Shared Intentional Engagement through Language and Phenomenal Experience. *Frontiers in Psychology, 5*. https://doi.org/10.3389/fpsyg.2014.01016

Feldman, M. J., Ma, R., & Lindquist, K. A. (2024). The Role of Interoception in Emotion and Social Cognition. In B J. Murphy & R. Brewer (Eds.), *Interoception* (pp. 125–149). Springer International Publishing. https://doi.org/10.1007/978-3-031-68521-7_5

Feuerriegel, S., Maarouf, A., Bär, D., Geissler, D., Schweisthal, J., Pröllochs, N., Robertson, C. E., Rathje, S., Hartmann, J., Mohammad, S. M., Netzer, O., Siegel, A. A., Plank, B., & Van Bavel, J. J. (2025). Using natural language processing to analyse text data in behavioural science. *Nature Reviews Psychology, 4*(2), 96–111. https://doi.org/10.1038/s44159-024-00392-z

Gamma, A., & Metzinger, T. (2021). The Minimal Phenomenal Experience Questionnaire (MPE-92M): Towards a Phenomenological Profile of "Pure Awareness" Experiences in Meditators. *PLOS ONE, 16*(7), e0253694. https://doi.org/10.1371/journal.pone.0253694

Gentzkow, M., Kelly, B., & Taddy, M. (2019). Text as Data. *Journal of Economic Literature, 57*(3), 535–574. https://doi.org/10.1257/jel.20181020

Gibbs, R. W., Costa Lima, P. L., & Francozo, E. (2004). Metaphor is Grounded in Embodied Experience. *Journal of Pragmatics, 36*(7), 1189–1210. https://doi.org/10.1016/j.pragma.2003.10.009

Hörberg, T., Larsson, M., & Olofsson, J. K. (2022). The Semantic Organization of the English Odor Vocabulary. *Cognitive Science, 46*(11), e13205. https://doi.org/10.1111/cogs.13205

Jraissati, Y., & Deroy, O. (2021). Categorizing Smells: A Localist Approach. *Cognitive Science, 45*(1), e12930. https://doi.org/10.1111/cogs.12930

Kraska-Szlenk, I. (2023). Embodied Lexicon: Body Part Terms in Conceptualization, Language Structure and Discourse. In B B. Lewandowska-Tomaszczyk & M.

Trojszczak (Eds.), *Language in Educational and Cultural Perspectives* (pp. 177–198). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-38778-4_9

Majid, A., & Burenhult, N. (2014). Odors are expressible in language, as long as you speak the right language. *Cognition, 130*(2), 266–270. https://doi.org/10.1016/j.cognition.2013.11.004

Mariko, M. L. (2025). Comparative Analysis of Lexical Density, Lexical Diversity, and Multiword Expressions in Russian, English, and French Legal Texts: Implications for Readability and Understandability. *Russian Linguistic Bulletin, 4*. https://doi.org/10.60797/RULB.2025.67.5

Martina, G. (2023). How we Talk about Smells. *Mind & Language, 38*(4), 1041–1058. https://doi.org/10.1111/mila.12440

Mudrik, L., Boly, M., Dehaene, S., Fleming, S. M., Lamme, V., Seth, A., & Melloni, L. (2025). Unpacking the Complexities of Consciousness: Theories and Reflections. *Neuroscience & Biobehavioral Reviews, 170*, 106053. https://doi.org/10.1016/j.neubiorev.2025.106053

Muñoz-Ortiz, A., Gómez-Rodríguez, C., & Vilares, D. (2024). Contrasting Linguistic Patterns in Human and LLM-Generated News Text. *Artificial Intelligence Review, 57*(10), 265. https://doi.org/10.1007/s10462-024-10903-2

Murphy, J., & Brewer, R. (2024). *Interoception: A Comprehensive Guide*. Springer International Publishing. https://doi.org/10.1007/978-3-031-68521-7

Nagornaya, A. V. (2014). *Discourse of the Inexpressible: Verbalism of Intra-Body Sensations*. Lenand.

Peirce, C. S., & Peirce, C. S. (1978). *Scientific Metaphysics* (C. Hartshorne & P. Weiss, Eds.). Belknap Press of Harvard Univ. Press.

Seth, A. K., & Tsakiris, M. (2018). Being a Beast Machine: The Somatic Basis of Selfhood. *Trends in Cognitive Sciences, 22*(11), 969–981. https://doi.org/10.1016/j.tics.2018.08.008

Shaules, J. (2020). *Language, Culture, and the Embodied Mind: A Developmental Model of Linguaculture Learning*. Springer Singapore.

Winter, B. (2019). *Sensory Linguistics: Language, Perception and Metaphor* (vol. 20). John Benjamins Publishing Company. https://doi.org/10.1075/celcr.20

Young, B. D. (2016). Smelling Matter. *Philosophical Psychology, 29*(4), 520–534. https://doi.org/10.1080/09515089.2015.1126814

Yu, N. (2020). *Linguistic Embodiment in Linguistic Experience: A Corpus-based Study* (Vol. 12). John Benjamins Publishing Company. https://doi.org/10.1075/clscc.12.c01yu

Zhang, J., Sun, M., Feng, Y., & Li, P. (2020). Learning Interpretable Relationships between Entities, Relations and Concepts via Bayesian Structure Learning on Open Domain Facts. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.717

Zhang, Y. (2011). Embodied Mind and Cross-cultural Narrative Patterns. In M. Callies, W. R. Keller & A. Lohöfer (Eds.), *Bi-Directionality in the Cognitive Sciences:*

Avenues, Challenges, and Limitations (pp. 171–180). John Benjamins. https://doi.org/10.1075/hcp.30.11zha

Zhao, X., Zheng, Y., & Zhao, X. (2023). Global bibliometric analysis of conceptual metaphor research over the recent two decades. *Frontiers in Psychology, 14*. https://doi.org/10.3389/fpsyg.2023.1042121

СВЕДЕНИЯ ОБ АВТОРАХ / THE AUTHORS

Барышников Павел Николаевич
pnbaryshnikov@pgu.ru
ORCID 0000-0002-0729-6698

Pavel Baryshnikov
pnbaryshnikov@pgu.ru
ORCID 0000-0002-0729-6698

Велис Лолита Андреевна
lolitavelis@yandex.com
ORCID 0009-0001-0881-1073

Lolita Velis
lolitavelis@yandex.com
ORCID 0009-0001-0881-1073

Атакуев Магомет Назирович
atakuevmagomet@gmail.com
ORCID 0000-0003-3135-4381

Magomet Atakuev
atakuevmagomet@gmail.com
ORCID 0000-0003-3135-4381

Статья поступила 22 мая 2025
одобрена после рецензирования 27 октября 2025
принята к публикации 15 декабря 2025

Received: 22 May 2025
Revised: 27 October 2025
Accepted: 15 December 2025