Мария Левченко

«А внутре у ней не(йр)онка»

СТАТИЧЕСКИЕ ВЕКТОРНЫЕ МОДЕЛИ В АНАЛИЗЕ ПОЭТИЧЕСКИХ ТЕКСТОВ

18+

Maria Levchenko

«Neural Inside»: Static Neural Models in Poetic Text Analysis

Мария Левченко
Болонский университет, Отделение классики и итальянистики, исследователь; кандидат филологических наук
marylevchenko@gmail.com.

Maria Levchenko
University of Bologna, Department of Classical Philology and Italian Studies, researcher; PhD
marylevchenko@gmail.com.

Ключевые слова: векторная модель, вычислительное литературоведение, семантический анализ, наивная поэзия, дистрибутивная семантика, векторная семантика

Keywords: computational literary studies, distributive semantics, word embeddings, semantic analysis, naive poetry, Russian literature

УДК: 81'322.2:821.161.1-1:004.85
DOI: 10.53953/08696365_2026_198_2_174

UDC: 81'322.2:821.161.1-1:004.85
DOI: 10.53953/08696365_2026_198_2_174

В данной статье исследуется применение векторных представлений слов для анализа различий между наивной и канонической поэзией. На материале сайта «Стихи.Ру» и русской поэзии XX века мы предлагаем систематический метод сравнения векторных моделей на основе анализа семантических сдвигов, дифференциации кластеров и стабильности соседей. Этот вычислительный подход предоставляет количественные метрики для комплексного описания фундаментальных различий в концептуализации ключевых поэтических понятий двух поэтических корпусов.

This article explores the application of word embeddings for analysing differences between naive and canonical poetry. Using corpora from *Stihi.Ru* and canonical poetry, we propose a systematic method for comparing vector models by analysing semantic shifts, cluster differentiation, and neighbour stability. Quantitative analysis suggests that naive poetry is manifested by denser semantic clusters, predictable associations, and concrete meanings, while canonical poetry operates with diffuse, multidimensional fields. This computational approach provides measurable evidence for fundamental differences in the conceptualisation of key poetic concepts.

Образованный читатель обычно легко отличает «подлинную» поэзию от любительского онлайн-творчества. Эта интуиция — довлатовское «прочтите строчки три» — опирается на предыдущий опыт и трудноуловимые сигналы текста, но итоговая оценка может радикально меняться в зависимости от внетекстовых факторов:

> Мне однажды приносят стихотворение, говорят: «Почитай». Неплохое акмеистическое стихотворение, может, известного мне акмеиста, а может, нет. Объясняют: «Нет, это написал наш сосед». Я говорю: «Тогда это абсолютно неинтересное стихотворение». — «Но он специально так пишет». — «Тогда интересно»[1].

Известный пример Дмитрия Пригова демонстрирует, что один и тот же текст воспринимается как тривиальный или значимый в зависимости от «имиджа

1 Цит. по: *Зорин А.Л.* Пригов как Пушкин // Театр. 1993. № 1. С. 117.

поэта», нашего знания о контексте и об авторской интенции. Ранние опыты канонического поэта могут быть неотличимы от наивного аутсайдерского письма, а сознательный минус-прием авангардиста — от непреднамеренной корявости дилетанта. Тончайшая грань между «докультурным» творчеством (до усвоения общих литературных конвенций) и «посткультурным» их нарушением (после усвоения и сознательного отказа), кажется, не поддается чисто имманентному анализу.

Однако алгоритмическое решение в этом случае не уступает читателю-эксперту. Так, простейшая логистическая регрессия, обученная на векторных представлениях (*OpenAI embeddings*) всего для 1000 текстов, с точностью в 99,5% определяет, относится ли ранее не встречавшееся ей стихотворение к «высокому» поэтическому канону или к «наивному» корпусу портала «Стихи.Ру». Впечатляющий результат (особенно если учесть, что классифицирующий алгоритм не «знает» ни года написания текста, ни авторской интенции: «специально» ли автор так пишет), который показывает, что различия между высокой и наивной поэзией объективно укоренены в самой семантической ткани текста и поддаются измерению.

Однако обладают ли векторные модели объяснительной, интерпретативной силой? Можем ли мы использовать их для полноценного анализа литературных текстов — не просто чтобы констатировать статистический факт, а эксплицировать те свойства, которые делают текст наивным или каноническим? Иными словами, можем ли мы операционализировать таким образом литературность поверх или за пределами чистых чисел?

* * *

Основа искусственного интеллекта, способного обрабатывать и генерировать тексты на человеческом языке, — векторные представления слов (*word embeddings*), абстрагированная семантическая информация, извлеченная алгоритмом из взаимных отношений слов в предложенных ему текстах. Каждое слово сохраняется нейронной моделью как *n*-мерный вектор, и в процессе ее обучения постепенно собирается многомерное векторное пространство, отражающее сложность выстроенной алгоритмом картины мира: каждый вектор соотнесен с соответствующим ему словом, каждое измерение вектора отражает синтаксические или семантические связи слова.

Векторные представления обеспечивают генерацию текстов с помощью больших языковых моделей (LLM), где в вектора переводятся не отдельные слова, а тексты, но они могут служить и эффективным аналитическим инструментом. Лежащая в основе архитектуры векторных моделей лингвистическая дистрибутивная гипотеза, формировавшаяся в 1950–60-е годы, исходит из того, что слова, встречающиеся в схожих контекстах, могут иметь близкое значение: «Скажи мне, кто твои соседи, и я скажу, кто ты»[2]. Опираясь на это представление о контекстуальной природе значения, алгоритмы машинного обучения агрегируют данные о контекстах каждого слова и переводят лингвистические закономерности в числовую форму, тем самым конструируя се-

2 «You shall know a word by the company it keeps» (*Firth J.R.* A Synopsis of Linguistic Theory, 1930–1955 // Studies in Linguistic Analysis. Oxford: Blackwell, 1957. P. 11).

мантическое пространство автоматически, без явно заданной онтологии или человеческого вмешательства, только на основе полученных на вход текстов. Этот подход к построению векторных представлений получил широкое распространение благодаря реализациям *word2vec*[3] и *GloVe*[4], пришедшим на смену более простым моделям, основанным на частотности слов в корпусе текстов.

Эксперименты с векторными моделями в лингвистике и далее в гуманитарных науках начались уже в 2014 году, практически сразу после презентации метода. Особенно продуктивным оказалось сравнение моделей, обученных на временных срезах больших текстовых корпусов, для выявления диахронических семантических сдвигов[5]. Интуитивная визуализация семантической эволюции слова *gay* из этого исследования, имитирующая движение слова в векторном пространстве от одних ближайших соседей (*cheerful*) в 1900 году к другим (*homosexual*) в 2005 году, стала классическим образцом того, какой результат может обеспечить подобный подход, а идея сопоставления векторных моделей была воспринята как перспективный метод для изучения семантических трансформаций[6].

---

3 *Mikolov T. et al.* Efficient Estimation of Word Representations in Vector Space // arXiv: 1301.3781. 2013 (URL: https://arxiv.org/abs/1301.3781); *Mikolov T., Sutskever I., Chen K., Corrado G., Dean J.* Distributed Representations of Words and Phrases and their Compositionality // Proceedings of the 27th International Conference on Neural Information Processing Systems. Vol. 2 / Ed. by C.J.C. Burges et al. Red Hook: Curran Associates, 2013. P. 3111–3119.

4 *Pennington J., Socher R., Manning C.* GloVe: Global Vectors for Word Representation // Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP) / Ed. by A. Moschitti, B. Pang, W. Daelemans. Doha: Association for Computational Linguistics, 2014. P. 1532–1543.

5 *Kulkarni V., Al-Rfou R., Perozzi B., Skiena S.* Statistically Significant Detection of Linguistic Change // Proceedings of the 24th International World Wide Web Conference / Ed. by A. Gangemi, S. Leonardi, A. Panconesi. Florence: ACM, 2015. P. 625–635.

6 На основе векторных моделей *word2vec* было сделано несколько исследований эволюции семантики на материале газетных корпусов (*Ros R., van Eijnatten J.* Disentangling a Trinity: A Digital Approach to Modernity, Civilization and Europe in Dutch Newspapers (1840–1990) // Book of Abstracts of DH2019. Utrecht: Utrecht University, 2019; *Hengchen S., Ros R., Marjanen J.* A Data-Driven Approach to the Changing Vocabulary of the «Nation» in English, Dutch, Swedish and Finnish Newspapers, 1750–1950 // Book of Abstracts of DH2019. Utrecht: Utrecht University, 2019; см. обзор в *Kutuzov A., Øvrelid L., Szymanski T., Velldal E.* Diachronic Word Embeddings and Semantic Shifts: a Survey // Proceedings of the 27th International Conference on Computational Linguistics / Ed. by E.M. Bender, L. Derczynski, P. Isabelle. Santa Fe: Association for Computational Linguistics, 2018. P. 1384–1397), по атрибуции авторства (на материале литературы на языке бенгали и англоязычной литературы: *Tripto N.I., Ali M.E.* The Word2vec Graph Model for Author Attribution and Genre Detection in Literary Analysis // arXiv:2310.16972. 2023 (URL: https://arxiv.org/abs/2310.16972), английского романа: *Eder M., Šeļa A.* One Word to Rule them All: Understanding Word Embeddings for Authorship Attribution // Digital Humanities 2022 Combined Abstracts. Tokyo: Alliance of Digital Humanities Organizations, 2022. P. 199–201). Интересно, что *word2vec*-модели могут быть использованы при построении графов персонажей (например, в качестве дополнительных признаков, существенно улучшающих результат при выстраивании графа по текстам Толкина: *Perri V., Qarkaxhija L., Zehe A., Hotho A., Scholtes I.* One Graph to Rule them All: Using NLP and Graph Neural Networks to Analyse Tolkien's Legendarium // Proceedings of the Computational Humanities Research Conference 2022. Vol. 3290. Antwerp: CEUR Workshop Proceedings, 2022. P. 291–317).

Изучение изменений в «ближайших соседях» (квази-синонимах) слова породило целый ряд исследований, построенных на сопоставлении векторных моделей, обученных на диахронических срезах одного корпуса текстов: подобная методика позволяет не только сравнивать состояние языка в двух временных точках, но и анализировать изменения как непрерывный процесс в течение последовательных временных интервалов. Так, например, в исследовании на материале голландского, шведского, финского и английского газетных корпусов XVIII–XX веков[7] проанализированы изменения семантизации концепта нации и национальной идентичности, сдвигавшегося от преимущественно экономических и военных ассоциаций в XVIII веке к увеличению фокуса на культурных аспектах, снижению доли экономической терминологии и его позднейшей институционализации. Эти более локальные изменения в квази-синонимах слов сигнализируют о «культурных», а не только языковых сдвигах[8], то есть позволяют обнаружить скрытые культурные смыслы и закономерности.

По сравнению с рекуррентными и тем более трансформерными архитектурами нейронных моделей, с их контекстуальными векторами больших размерностей (768 для обычных трансформеров типа BERT и 1536–3072 для векторов *OpenAI*), сегодня статические вектора *word2vec* размерности 50–500 могут показаться устаревшими. Последний обзор техник машинного обучения в вычислительном литературоведении (*computational literary studies*) показывает, что фокус внимания исследователей смещается в сторону трансформеров и их последней инкарнации LLM[9]. Однако статические модели сохраняют значительную эпистемологическую ценность: для выявления семантических отношений между словами они показывают лучшие[10] или сопоставимые[11] результаты.

Их интерпретируемость, вычислительная эффективность, применимость к малым корпусам текстов и способность улавливать семантические сдвиги отдельных слов делают их особенно подходящими для гуманитарных исследований, где выявление концептуальных отношений важнее генеративных возможностей. Архитектура *word2vec*, с ее способностью операционализировать культурные ассоциации через векторные операции, зачастую остается оптимальным решением для задач *digital humanities*.

В последние годы было сделано несколько интересных попыток применения векторных моделей к изучению литературных текстов. Работа Антона Эрмантраута и др. посвящена анализу различий/сходств между реалистической и модернистской поэзией[12], построенному на корпусе антологий немецкой поэ-

---

7 *Hengchen S., Ros R., Marjanen J., Tolonen M.* A Data-Driven Approach to Studying Changing Vocabularies in Historical Newspaper Collections // Digital Scholarship in the Humanities. 2021. Vol. 36. Issue Supplement 2. P. ii109–ii126.

8 *Hamilton W.L., Leskovec J., Jurafsky D.* Op. cit. P. 1499.

9 *Hatzel H.O., Stiemer H., Biemann C., Gius E.* Machine Learning in Computational Literary Studies // Information Technology. 2023. Vol. 65. № 4–5. P. 200–217.

10 *Ehrmanntraut A., Hagen T., Konle L., Jannidis F.* Type- and Token-Based Word Embeddings in the Digital Humanities // Proceedings of the Computational Humanities Research Conference 2021 (CHR 2021). Amsterdam; Antwerp: CEUR Workshop Proceedings, 2021. P. 16–38.

11 *Tripto N.I., Ali M.E.* Op. cit.

12 *Ehrmanntraut A., Hagen T., Jannidis F., Konle L., Kröncke M., Winko S.* Modeling and Measuring Short Text Similarities. On the Multi-Dimensional Differences between German Poetry of Realism and Modernism // Journal of Computational Literary Studies. 2022. Vol. 1. № 1. P. 1–30.

зии (3039 текстов для реализма и 2882 текста для модернизма). Исследование основано на измерении близости текстов по содержанию, форме, стилю и выраженным в тексте эмоциям. Результаты релятивизируют устоявшееся представление о резком переломе при переходе от реализма к модернизму: вместо четкой границы между направлениями авторы обнаруживают постепенную эволюцию и заключают, что традиционное противопоставление двух литературных систем может быть излишне категоричным. При этом анализ подтверждает, что средняя дистанция между текстами разных направлений (реализм — модернизм) в векторном пространстве действительно больше, чем средняя дистанция между текстами внутри реалистического корпуса, однако и модернизм сам по себе более неоднороден, чем реализм. Таким образом, сравнительный анализ нескольких векторных моделей позволяет не просто расширить спектр традиционных инструментов анализа литературных текстов, но и верифицировать устоявшиеся историко-литературные представления.

Наиболее очевидный и известный способ сопоставления двух моделей, обученных на разных срезах одного корпуса или на разных корпусах, — сравнение списка ближайших соседей для наиболее интересных нам слов из их общего словаря: «О сходстве или несходстве значений слов можно судить, в частности, по их квази-синонимам (в нашем случае вернее говорить об ассоциатах), то есть по словам, наиболее близким по контекстуальному окружению к данному»[13]. И хотя такой подход требует времени и интуиции исследователя, он остается одним из основных способов применения *word2vec*-моделей для анализа литературных текстов[14]. Однако потенциальные возможности векторных представлений для анализа литературных текстов гораздо шире и могут служить основой для системного анализа семантического пространства, построенного алгоритмом на основе дистрибуции всех значимых слов представительного корпуса текстов.

Перспективным может быть применение такого подхода к исследованию различий между литературными традициями, трудно поддающихся формализации через традиционное «медленное» чтение. Наивная и каноническая поэзия представляют собой именно такой случай: хотя квалифицированный читатель обычно без труда различает их, формализация и объективация этих интуитивных различий остается нетривиальной задачей. Если у нас есть векторные модели семантического пространства на основе текстов наивной и ка-

13 *Орехов Б.В.* Стихи и проза через призму дистрибутивной семантики // Острова любви БорФеда: сборник к 90-летию Бориса Федоровича Егорова. СПб.: Росток, 2016. С. 653.

14 Например, Б.В. Орехов анализирует различия между поэтическим и прозаическим корпусами НКРЯ с помощью обученных на этих корпусах *word2vec*-моделей и приходит к выводу о том, что семантика слов в прозаическом и поэтическом контекстах существенно отличается, а в одной из недавних работ, построенных на этом методе, векторная модель помогает при анализе различий в концептуализации слова *queer* у Вирджинии Вульф в контексте англоязычной прозы 1850–1990-х годов. Вычислив с помощью *word2vec*-модели 100 ближайших соседей, или квази-синонимов, *queer* у Вирджинии Вульф, автор сравнивает этот список с соответствующими у Джойса, Фицджеральда, Лоренса, Стайн и Мэнсфилд и демонстрирует, что у Вульф этот список гораздо «позитивнее», чем у ее современников, за исключением Джойса (*Shin H.* Analyzing the Positive Sentiment towards the Term «Queer» in Virginia Woolf through a Computational Approach and Close Reading // Journal of Computational Literary Studies. 2022. Vol. 1. № 1. P. 1–26).

нонической поэзии, можем ли мы, проанализировав и сравнив конструктивные особенности этих пространств, получить верифицируемые выводы об их различиях? Фиксируются ли векторной моделью особенности в семантической организации наивной и канонической поэзии?

Определяя в 2001 году «наивную литературу» в одноименном сборнике, М.Л. Лурье описывает не столько тексты наивной литературы, сколько их авторов, реальных или реконструируемых по текстам: это «письменные (литературные) тексты, созданные человеком, не владеющим нормами письменной (литературной) речи»[15], и дальше уточняет коллективный портрет: «дискурсивные дилетанты», которые «не являются квалифицированными потребителями литературной продукции, находящимися в одном социокультурном пространстве с ее производителями»[16]. Представленные же в сборнике тексты, каждый из которых прекрасен по-своему, тем не менее не совсем укладывались в единую картину, и Лурье сетовал:

> До сих пор ученые интерпретации, по сути, касались единичных фактов, а не явления в целом, общие же особенности явления в свою очередь не могут быть определены сколько-нибудь четко, пока в руках исследователей столь фрагментарный, разрозненный и по большей части случайный материал[17].

Проект «Стихи.Ру», запущенный в 2000 году (первый текст 1 марта 2000 года), насчитывает сегодня огромный корпус в 64 млн текстов[18], часто определяемых в исследовательской литературе как «наивная» или «интернет-поэзия». Он предоставил интернет-пользователям платформу для публикации их поэтических текстов, а исследователям наивной поэзии — с избытком материала для изучения. Кажется, он не вполне соответствует описанной Лурье специфике, по крайней мере, относительно изолированности наивных литераторов: «Стихи.Ру» образует свою субкультуру или целый спектр субкультур[19], и с тех пор, как сайт стал заметным явлением Рунета, уже нельзя говорить о том, что наивные сочинители не догадываются о существовании друг друга, как деревенские наивные писатели[20]. Использование цифровой платформы не предполагает ни полной изоляции от литературного процесса, ни полной интеграции в сложившиеся литературные институции. Скорее, это способ альтернативной литературной социализации, причиной выбора которого не обязательно является некомпетентность авторов. Тем не менее сохраняется представление о текстах «Стихов.Ру» как о наивной поэзии — с ее упрощенной поэтикой и

---

15 *Лурье М.Л.* О феномене наивного сочинительства // «Наивная литература»: исследования и тексты / Сост. С.Ю. Неклюдов. М.: Московский общественный научный фонд, 2001. С. 18.

16 Там же. С. 20.

17 Там же. С. 20–21.

18 Этот огромный корпус текстов не вполне чист от «чужеродных» примесей: страницы известных поэтов, прозаические вкрапления, первые публикации поэтов, вышедших потом в пространство «настоящей» литературы: считается, что «некоторые известные ныне поэты нового поколения начинали именно на “стихире”. Авторы “стихиры” впоследствии образовали ядро двух крупных литобъединений — московского “Рукомоса” и питерского “Пиитера”» (*Галина М.* Поэзия онлайн // Звезда. 2007. № 2 (URL: https://magazines.gorky.media/znamia/2007/2/mariya-galina-2.html)).

19 *Анистратенко А.* Лоскутное одеяло Стихиры // Сетевая словесность (URL: https://www.netslova.ru/anistratenko/stihira.html).

20 *Лурье М.Л.* Указ. соч. С. 22.

особыми отношениями с литературной традицией и языком. Это представление, однако, можно было бы эмпирически верифицировать через систематический анализ текстовых особенностей. Если мы действительно имеем дело с особым типом бытования поэзии, пусть и в контексте цифровой платформы, то должны существовать объективные, измеримые различия между ним и канонической поэзией — различия, которые можно выявить с помощью современных методов компьютерного анализа.

Огромный объем текстовых данных провоцирует применение методов *distant reading*[21], и на этом материале уже проводились количественные исследования. В 2013 году А. Бонч-Осмоловская и Б. Орехов исследовали лексику наивной и высокой поэзии, опираясь на списки наиболее частотных слов у авторов «Стихи.Ру» и поэтического подкорпуса НКРЯ. Результаты показывают различия в частоте местоимений и тематической лексики: в отличие от высокой поэзии в наивной акцент смещен на бытовую лексику и «простые» эмоциональные концепты, что характерно для современной разговорной речи[22]. Исследуя темы Крыма в русскоязычной наивной поэзии на материале текстов «Стихи.Ру», Р. Лейбов и Б. Орехов использовали тематическое моделирование и определяли ключевые темы через анализ часто встречающихся словосочетаний, и пришли к выводу, что наивные поэты воспроизводят пропагандистские клише без особого влияния «высокой» поэзии, сохраняют простую стиховую форму, но существенно расходятся с литературной традицией в содержании: «…на тематическом уровне мы не прослеживаем отчетливого влияния “высокой” поэзии на авторов Stihi.ru»[23]:

> Парадоксальным образом традиционалистская по форме (стиховое членение, ритм, рифма) реакция наивных поэтов резко порывает с литературной традицией содержательного воплощения такой реакции. <…> У наивных авторов балансирование смещается в сторону формы, а при создании плана содержания наивная литература оказывается начисто лишена литературной (и именно литературной) памяти[24].

Такой тип взаимодействия с традицией близок интертекстуальным стратегиям предшествующей волны «дискурсивных дилетантов» — поэтов массовых отделений Пролеткульта, заимствовавших из современной им «настоящей» поэзии только форму[25]. С.Ю. Неклюдов предполагает, что «моделью для наивной поэзии (особенно крестьянской) является скорее русская классика (Пушкин, Лермонтов, Некрасов, Кольцов, Никитин и др.), тогда как проза тяготеет к более современным автору литературным образцам»[26]. Для авторов Пролеткуль-

---

21 *Моретти Ф.* Дальнее чтение / Пер. с англ. А. Вдовина, О. Собчука, А. Шели; науч. ред. перевода И. Кушнарева. М.: Издательство Института Гайдара, 2016.

22 *Бонч-Осмоловская А.А., Орехов Б.В.* Некоторые применения корпусных методов к наивной поэзии // Статьи на случай: Сборник в честь 50-летия Р.Г. Лейбова. М.: Объединенное гуманитарное издательство Ruthenia.ru, 2013. С. 22–36.

23 *Лейбов Р., Орехов Б.В.* Между политикой и поэтикой: топика Крыма в современной русскоязычной наивной лирике // Шаги / Steps. 2022. Т. 8. № 2. С. 229.

24 Там же. С. 230.

25 *Левченко М.А.* Индустриальная свирель: Поэзия Пролеткульта 1917–1921 гг. СПб.: СПГУТД, 2007. С. 119.

26 *Неклюдов С.Ю.* От составителя // «Наивная литература»: исследования и тексты / Сост. С.Ю. Неклюдов. М.: Московский общественный научный фонд, 2001. С. 10.

та (за исключением нескольких ярких центральных фигур, активно черпавших образы из современной им модернистской поэзии) ориентиром был школьный канон. Предположительно, для авторов «Стихи.Ру» будет актуальнее корпус русской поэзии начала XX века[27], ставший каноническим к началу XXI века, за счет «отставания по фазе по отношению к “основному” национальному литературному процессу», о котором пишет С.Ю. Неклюдов.

Размышляя о наивной поэзии, мы неизбежно сравниваем ее не столько с актуальными поэтическими тенденциями, сколько с контекстом предшествующей «канонизированной» литературы, который и для самих наивных авторов, кажется, служит важным ориентиром: актуальные поэты чаще всего не попадают в зону их внимания или сферу культурной компетенции. Проверить теоретические предположения о различиях между канонической и наивной поэзией можно через количественный анализ семантических структур с помощью векторных представлений *word2vec*, сопоставив два репрезентативных поэтических корпуса.

Для обучения векторной модели наивной поэзии был использован полнотекстовый корпус «Стихов.Ру» за 2002 год (226 575 текстов, 2 095 415 словоупотреблений, после лемматизации 70 829 уникальных слов). Для сравнения в нашем распоряжении есть поэтический корпус русской поэзии второй половины — первой половины XX века[28], содержащий 31 639 текстов (количество словоупотреблений 2 288 350, после лемматизации 79 247 уникальных слов в словаре). Перед обучением тексты были разделены на предложения, а потом приведены к нормальной форме с помощью пакета *mystem* от *Yandex*. Параметры обучения использовались общие для обоих корпусов: размерность вектора 100, окно 5, 100 эпох обучения, кроме того, отсекались слова, встречающиеся в корпусе менее 10 раз. Лемматизация обеспечила статистически надежное пересечение словарей: после обучения моделей их общий словарь составил 15 579 слов, что было бы недостижимо при работе со словоформами.

Базовые метрики. Первичная проверка моделей после выравнивания векторных пространств ортогональным преобразованием Прокруста[29] позво-

---

27 В пользу такого предположения свидетельствует, например, тот факт, что линейный классификатор на векторных представлениях из модели *OpenAI*, о котором шла речь в начале статьи, среди нескольких допущенных им ошибок отнес стихотворение Ахматовой «Я улыбаться перестала...» к наивной поэзии.

28 Неполный список авторов, чьи тексты вошли в корпус: Г. Адамович, И. Анненский, А. Ахматова, Э. Багрицкий, К. Бальмонт, А. Белый, А. Блок, В. Брюсов, И. Бунин, М. Волошин, З. Гиппиус, А. Григорьев, Н. Гумилев, С. Есенин, Н. Заболоцкий, Вяч. Иванов, Г. Иванов, Н. Клюев, М. Кузмин, О. Мандельштам, В. Маяковский, В. Набоков, Б. Пастернак, И. Северянин, Ф. Сологуб, Ф. Тютчев, А. Фет. Около 75% текстов корпуса относятся к периоду 1900–1940-х годов, поэтому хронологически он более релевантен для сравнения с наивными текстами «Стихов.Ру»: поэтический подкорпус НКРЯ содержит тексты XVIII–XIX веков, вносящие существенный шум в лексический состав. Впрочем, хотя временной зазор между корпусами все равно сказывается при сравнении моделей, идентифицировать обусловленные этим зазором расхождения очень легко, что мы увидим ниже.

29 Если вектора не конструируются последовательно на диахронических срезах однородного корпуса, а представляют разные корпуса текстов, необходимы дополнительные (математические) операции для «выравнивания» векторного пространства одной модели относительно другой (других). Стандартный подход для этого — ортогональное преобразование Прокруста (*Schönemann P.H.* A Generalized Solution of the Orthogonal Procrustes Problem // Psychometrika. 1966. Vol. 31. P. 1–10), математический

ляет нам сравнить предсказуемость и стабильность семантических ассоциаций в обоих корпусах с помощью количественных метрик[30]. Более высокая средняя степень близости между квази-синонимами в наивной поэзии (0,5245 против 0,4421) указывает на ее тяготение к устойчивым, конвенциональным семантическим связям[31]. Значительно более высокая стабильность соседства (0,5568 для наивной поэзии против 0,2722 для канонической) свидетельствует о предсказуемости и однородности ассоциаций. Большая плотность семантических сетей (0,3931 для наивной поэзии против 0,2737 для канонической) демонстрирует тенденцию к формированию компактных, четко дифференцированных смысловых полей с сильными внутренними связями и более выраженными границами между кластерами. Наконец, меньшая вариативность в баллах сходства (отношение 0,6292) отражает сравнительную ограниченность семантического репертуара в наивной поэзии. В совокупности эти показатели подтверждают гипотезу о том, что наивная поэзия тяготеет к буквальному, кон-

---

метод преобразования/разворачивания одного векторного пространства для лучшего совмещения с другим векторным пространством, при котором сохраняются относительные расстояния между словами для каждой модели (*Kutuzov A., Øvrelid L., Szymanski T., Velldal E.* Diachronic Word Embeddings and Semantic Shifts: a Survey // Proceedings of the 27th International Conference on Computational Linguistics / Ed. by E.M. Bender, L. Derczynski, P. Isabelle. Santa Fe: Association for Computational Linguistics, 2018. P. 1389), более подробный список способов подготовки двух моделей для сравнительного анализа см. в: *Fomin V., Bakshandaeva D., Rodina J., Kutuzov A.* Tracing Cultural Diachronic Semantic Shifts in Russian Using Word Embeddings: Test Sets and Baselines // arXiv:1905.06837. 2019 (URL: https://arxiv.org/abs/1905.06837).

30 Стабильность соседства рассчитывается как пересечение множеств ближайших соседей первого и второго порядка: $S(w) = |N_1(w) \cap N_2(w)| / |N_1(w)|$, где N_1 — k ближайших соседей слова w, N_2 — объединение k ближайших соседей каждого из N_1. Плотность семантической сети вычисляется как средняя попарная косинусная близость между всеми соседями: $D(w) = (2/k(k-1)) \times \Sigma_{i<j} \cos(n_i, n_j)$. Вариативность оценивается через стандартное отклонение косинусных сходств и коэффициент вариации ($CV = \sigma/\mu$).

31 Имеет смысл уточнить, почему в нашем случае минимизируется отмеченное в литературе влияние стохастических вариаций на стабильность *word2vec*-моделей (*Antoniak M., Mimno D.* Evaluating the Stability of Embedding-Based Word Similarities // Transactions of the Association for Computational Linguistics. 2018. Vol. 6. P. 107–119; *Wendlandt L., Kummerfeld J.K., Mihalcea R.* Factors Influencing the Surprising Instability of Word Embeddings // Proceedings of NAACL-HLT / Ed. by M.A. Walker, H. Ji, A. Stent. New Orleans: Association for Computational Linguistics, 2018. P. 2092–2102; *Маслинский К.А.* Сто лет счастья в детской литературе (1920–2020): сталинский канон и его долгосрочные последствия // Шаги / Steps. 2022. Т. 8. № 4. С. 226–247). Во-первых, наблюдаемые различия между моделями (например, стабильность соседства 0,5568 vs 0,2722 — более чем в два раза) значительно превышают диапазоны нестабильности, отмеченные в упомянутых работах для высокочастотных слов (5–15%). Во-вторых, использование 100 эпох обучения, фокус на пересекающемся словаре высокочастотных слов и применение ортогонального преобразования Прокруста снижают влияние случайных факторов. Преобразование Прокруста специально направлено на устранение вариаций, связанных с различными инициализациями векторных пространств, что является стандартной практикой для межкорпусных сравнений (*Hamilton W.L., Leskovec J., Jurafsky D.* Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change // Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) / Ed. by K. Erk, N.A. Smith. Berlin: Association for Computational Linguistics, 2016. P. 1489–1501). В-третьих, различия наблюдаются по множественным независимым метрикам (плотность семантических сетей, кластерная структура, индексы качества кластеризации), что указывает на системный характер различий, а не на артефакты обучения.

кретному использованию языка с предсказуемыми ассоциациями, в то время как каноническая поэзия характеризуется многомерными, диффузными семантическими полями, обеспечивающими ее метафорическую сложность и широту ассоциативных связей.

Семантические кластеры. Кластерный анализ векторных представлений словарей обеих моделей позволяет описать скрытую концептуальную структуру каждой традиции — то, как слова естественным образом группируются в семантическом пространстве на основе закономерностей их использования. После тестирования различных вариантов (от 10 до 50 кластеров) с использованием жесткой кластеризации *K-Means*[32] была выбрана оптимальная конфигурация с 25 кластерами, дающая наилучший баланс между детальностью и интерпретируемостью (наивысшее значение нормализованной взаимной информации 0.2711).

Для оценки качества кластеризации использовались две взаимодополняющие метрики: индекс Калинского-Харабаша и коэффициент силуэта[33]. Наблюдаемые различия в показателях — значительно более высокий индекс Калинского-Харабаша (74.87 против 50.23) и коэффициент силуэта (-0.05 против -0.1) для наивной поэзии — согласуются с гипотезой о различной организации семантических пространств. Эти результаты можно интерпретировать как свидетельство более четких границ между семантическими кластерами в наивной поэзии, в то время как каноническая поэзия демонстрирует тенденцию к более диффузной организации, где концепты менее жестко разграничены, что согласуется с традиционными литературоведческими представлениями о метафоричности канонической поэзии и конвенциональности наивного поэтического языка.

Кластеризация также дает возможность визуально представить организацию семантических полей в обеих моделях (см. Илл. 1)[34]. Примечательно, что наибольшее сходство между моделями (совпадение элементов около 50%) наблюдается в кластерах, относящихся к физическому миру и сенсорному восприятию:

- *пение, звучать, звук, голос, слышать* (каноническая) / *звук, слышаться, слышный, звонкий, пение* (наивная);
- *берег, море, волна, вода, корабль* / *берег, лодка, корабль, океан, море*;

---

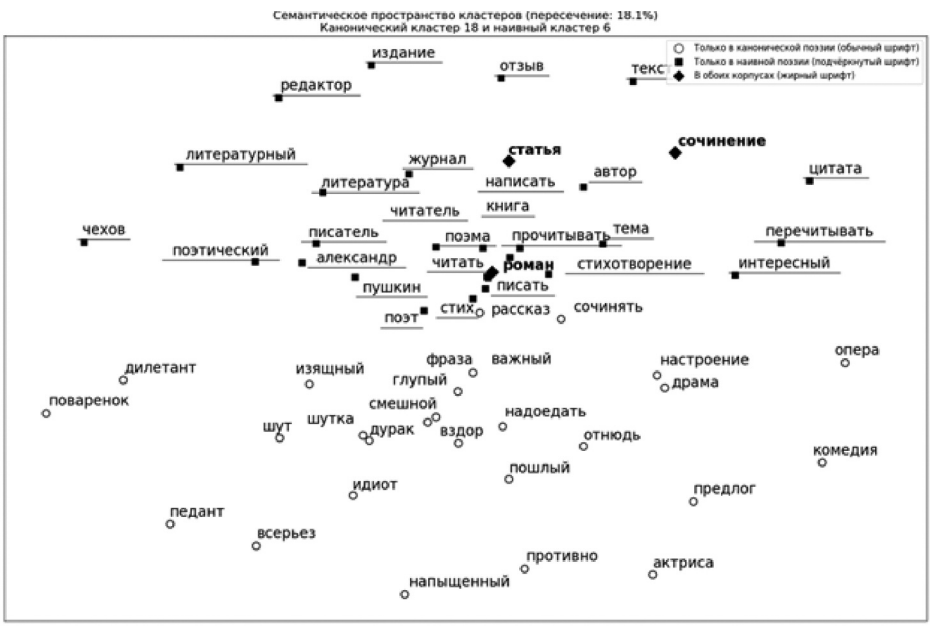
32 Использовался алгоритм *K-Means* из библиотеки *scikit-learn*. Оптимальное число кластеров определялось методом локтя (*elbow method*) с автоматическим поиском точки перегиба через *KneeLocator*.

33 Индекс Калинского-Харабаша оценивает разделимость кластеров: более высокий показатель свидетельствует о том, что кластеры более четко очерчены и лучше отделены друг от друга. Коэффициент силуэта измеряет степень сходства объекта с другими объектами своего кластера по сравнению с объектами других кластеров. Отрицательное значение силуэта, свидетельствующее о том, что семантические кластеры естественным образом пересекаются, типично для языковых данных, тем более что мы используем вектора *word2vec*, которые создают единый вектор слова вне зависимости от дифференцируемости его возможных контекстов.

34 Для человеческого восприятия недоступны многомерные объекты, какими являются векторы слов в *word2vec*-модели, располагающиеся в 50–300 измерениях, поэтому обычно производят снижение размерности векторов до 2–3-мерного пространства с помощью стандартных техник (*PCA*, *t-SNE* и др.). Такое редуцированное представление, сохраняя относительные дистанции между векторами и их взаимное расположение, позволяет визуализировать получившееся векторное пространство с очевидным упрощением отношений даже на бумаге. Однако при всех количественных расчетах мы продолжаем работать с полноразмерными векторами.

- *чай, пить, пиво, хлеб, водка / вкусный, салат, ложка, сало, суп*;
- *душистый, цветок, цвести, роза, цвет / трава, душистый, зелень, ягода, сирень.*

Однако за пределами универсальных физических концептов модели демонстрируют существенное несовпадение семантических структур. Особенно показателен анализ металитературных концептов. В «наивной» модели (Илл. 1) четко выделяется кластер, включающий термины поэтического творчества (*стихотворение, стих, писатель, читать, автор, написать, поэма, текст, поэт, поэзия*), связанный с легитимирующими литературными формами и институциями (*книга, журнал, статья, издание, текст*) и процессуальными глаголами литературного производства (*писать, читать, перечитывать*). В канонической же поэзии соответствующий сегмент семантического пространства (пересечение только на словах *сочинение, статья, роман*) насыщен ироническими коннотациями (*шут, напыщенный, дурак, смешной, глупый, важный, пошлый*). Это наблюдение позволяет предположить, что более прямая и сфокусированная саморефлексия у наивных поэтов направлена на институциональную самореализацию, что согласуется с отмеченной М.Л. Лурье «специфической амбициозностью» наивных литераторов[35]: авторы платформы ориентируются на внешние формы литературного быта, адаптированные и оцифрованные механиками литературной социальной сети (конкурсы, рейтинги, рецензии, альманахи), которые только и валидируют «поэтическое» в этом виртуальном контексте.



Илл. 1. Наложение семантических кластеров в наивной и канонической поэзии: «металитературный» кластер

35 *Лурье М.Л.* Указ. соч. С. 23.

Семантический сдвиг. Хотя кластерный анализ выявляет принципиальные различия в организации семантических полей, для детального понимания их своеобразия необходимо более точно оценить семантические сдвиги между моделями. Полноценное сравнение двух корпусов может быть достигнуто через комплексный анализ двух базовых параметров векторных представлений: смещения векторов одних и тех же слов в одной модели относительно другой и различия/сходства в составе их ближайших соседей. Анализируя спектр семантических смещений от наиболее сильного до слабых вариаций, мы можем выявить ключевые расхождения в стратегиях семантизации между канонической и наивной поэзией. На основе количественных метрик (косинусная близость между векторами слова в двух моделях и пересечение топ-10 ближайших соседей для каждого из слов) мы выделили четыре типа семантического смещения двух моделей относительно друг друга, что позволяет систематизировать наблюдаемые различия:

1. Зона сильного смещения (сильное смещение векторов одновременно с изменением в ближайших соседях слов: косинусная близость <0.8, пересечение соседей <0.3[36]). Наиболее частотные слова из этой зоны: *душа, сердце, время, хотеть, каждый*.

2. Зона семантического сдвига (существенное смещение векторов при сохранении той же группы квази-синонимов: косинусная близость <0.8, пересечение соседей ≥0.3; слова подвержены менее значительному изменению смысла, но весь смысловой блок одной модели (вместе с таргетным словом) сдвинут относительно другой: *идти, видеть, белый, счастье, мечта, лицо*.

3. Зона контекстуального сдвига (стабильные вектора при новых соседях: косинусная близость ≥0.8, пересечение соседей <0.3): *любовь, мир, жить, человек, земля, солнце, слово*.

4. Стабильная зона (без существенных изменений, стабильные вектора и схожий набор ближайших соседей: косинусная близость ≥0.8, пересечение соседей ≥0.3): *день, жизнь, глаз, рука, ночь, свет, сон, небо*.

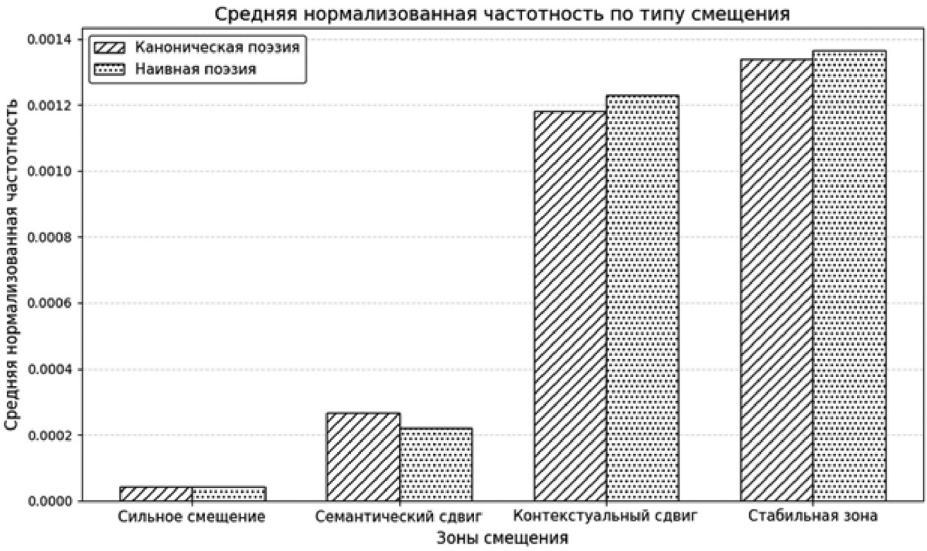
Наиболее частотные слова в обоих корпусах одновременно наиболее близки в обеих моделях как по семантическому вектору («ночь»: близость векторов между моделями 0.8652, «день»: 0.8673, «земля»: 0.8687, «вода»: 0.8697, «сон»: 0.8853, «окно»: 0.8873, «луч»: 0.8875, «небо»: 0.8916, «кровь»: 0.8928, «год»: 0.8970, «петь»: 0.9017), так и по своим ближайшим соседям, что соответствует и лингвистическому представлению об эволюции языка: более частотные слова меньше подвержены изменениям[37]. Как видно на графике (Илл. 2), средняя нормализованная частотность «стабильных» слов в 25 раз выше, чем у слов с силь-

---

36 Пороговые значения были установлены следующим образом: cos = 0.8 соответствует углу ~37° между векторами, что является стандартным порогом высокого семантического сходства в литературе по дистрибутивной семантике (*Turney P.D., Pantel P.* From Frequency to Meaning: Vector Space Models of Semantics // Journal of Artificial Intelligence Research. 2010. Vol. 37. № 1. P. 141–188); overlap = 0.3 — эмпирически установленный порог, разделяющий слабое (<0.3) и существенное (≥0.3) пересечение семантических окрестностей.

37 Эта закономерность отражает фундаментальное свойство языковой эволюции (*Pagel M., Atkinson Q.D., Meade A.* Frequency of Word-Use Predicts Rates of Lexical Evolution throughout Indo-European History // Nature. 2007. Vol. 449. № 7163. P. 717–720; *Bybee J.L.* Frequency of Use and the Organization of Language. New York: Oxford University Press, 2007. P. 5–22), а не только алгоритмический эффект: мы наблюдаем стабильность между двумя независимыми корпусами разных поэтических традиций.

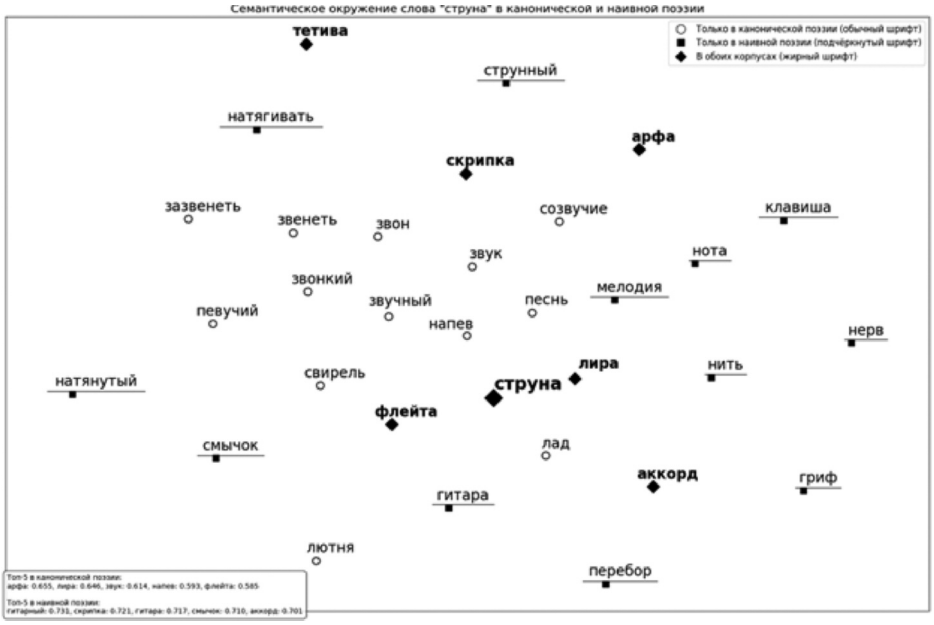
ным смещением, то есть базовые, наиболее частотные единицы поэтического языка сохраняют стабильность значения независимо от поэтической традиции. Различия же в корпусах раскрываются через менее частотные слова. Особый интерес в этом контексте представляют слова из зоны контекстуального сдвига. Они демонстрируют близкий уровень частотности к стабильным словам, но при этом обнаруживают значительные различия в контекстуальном окружении, что делает их особенно ценными для анализа: при стабильном базовом значении (обусловленном высокой частотностью) в наивной поэзии существенно меняются их контексты относительно поэзии канонической.



Илл. 2. Средняя нормализованная частотность слов по силе семантического сдвига между моделями

Получив данные о семантических кластерах и об основных зонах семантического сдвига, мы можем систематизировать сопоставление двух моделей в зависимости от того, что нас в данный момент интересует больше: изменения в одном или нескольких семантических полях (и тогда мы рассматриваем конкретные кластеры) или области наибольших различий между моделями. Попробуем проиллюстрировать последний вариант на нескольких примерах, выбранных на основе двух критериев: 1) слова из зон контекстуального и сильного семантического сдвига; 2) лексемы с высокой частотностью, позволяющие минимизировать влияние случайных флуктуаций в данных.

Струна (Илл. 3). Векторы слова «струна» в двух моделях демонстрируют умеренную близость (0.82) при малом пересечении ближайших соседей (0.17, общие соседи: *арфа, флейта, скрипка, аккорд, лира, тетива*). Анализ соседей (топ-20 для каждой модели) показывает существенные различия в контекстуальном окружении при стабильном базовом значении. В канонической поэзии доминируют разнообразные «звуковые» ассоциации (*звук, звон, звучный, напев, песнь, звенеть* и т.д.). В наивной поэзии к «струне» ближе музыкальные инструменты и их атрибуты (*струнный, клавиши, нота, гриф, перебор, гитара, смычок, натягивать*).



Илл. 3. Квази-синонимы для слова «струна» (зона контекстуального сдвига)

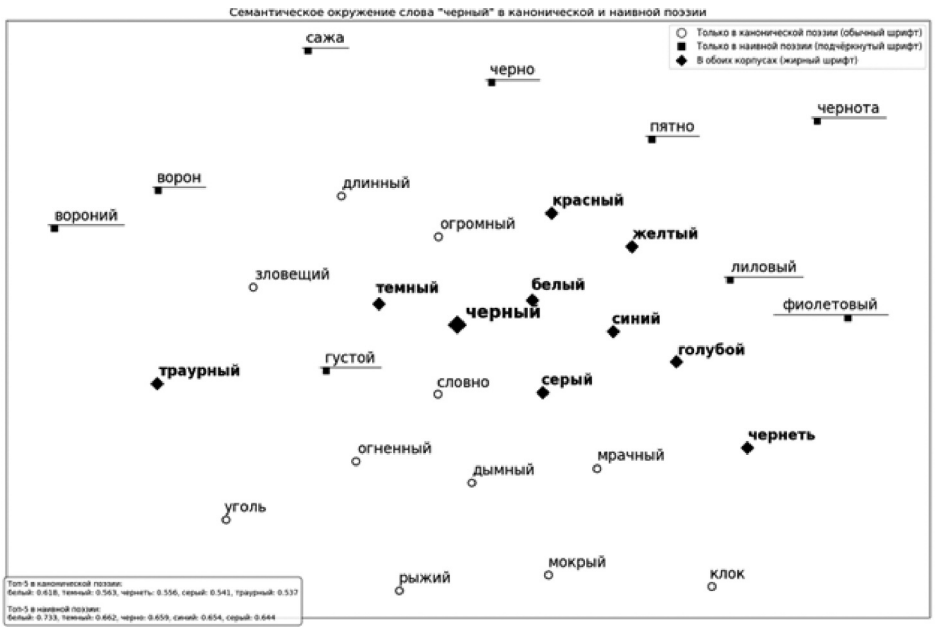
В семантическом окружении «***слова***» (Илл. 4, зона контекстуального сдвига, косинусная близость векторов 0.871) пересечение квази-синонимов составляет 0.25: *фраза, строка, речь, сказать, рифма, признание, словом, смысл*. Различия наблюдаются в уникальных для каждого корпуса ближайших соседях:

Илл. 4. Квази-синонимы для слова «слово» в канонической и наивной поэзии

в канонической поэзии *мысль, чувство, язык, стих, понимать, простой, понятный, звучать, незначащий*, в наивной поэзии: *буква, слог, строчка, словечко, клятва, обещание, значение, недосказанный*. Эти расхождения могут свидетельствовать о различных функциональных акцентах в концептуализации «слова»: в канонической поэзии — на интеллектуальном и герменевтическом аспекте, в наивной — на формальном и перформативном.

Векторы лексемы «***черный***» (Илл. 5, зона семантического сдвига) демонстрируют среднюю степень близости между моделями (близость векторов 0.7, пересечение соседей 0.33). Помимо общего для обеих моделей цветового ряда, в канонической поэзии ближайшие соседи дифференцируются по нескольким семантическим группам: визуальные характеристики (*огненный, сизый, дымный*), тактильные ассоциации (*мокрый, сырой*), эмоциональные коннотации (*зловещий, страшный, мрачный*) и символические значения (*траурный*).

В наивной поэзии продолжается цветовой ряд: *лиловый, фиолетовый, черно, чернота*, а кроме того, добавляются общеязыковые ассоциации: *сажа, уголь, ворон, вороний*.

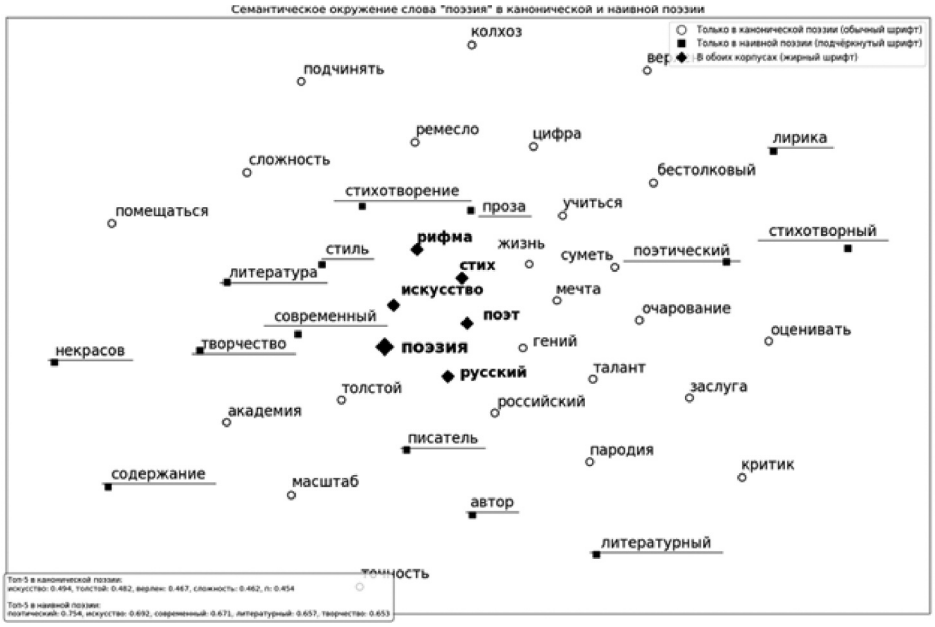


Илл. 5. Квази-синонимы для слова «черный» в канонической и наивной поэзии

Завершим наш сопоставительный анализ визуализацией семантического пространства лексемы «поэзия» (Илл. 6), демонстрирующего фундаментальные различия в саморефлексивности двух поэтических традиций. Количественные параметры подтверждают существенное расхождение между моделями: косинусное сходство векторов составляет 0.6, а пересечение ближайших соседей лишь 0.17. Общее ядро семантического поля формируют лексемы *рифма, стих, поэт, искусство и русский*, однако структурная организация этого поля принципиально различна.

В канонической поэзии наблюдается низкая концентрация семантических связей: максимальный коэффициент близости составляет всего 0.494 (*искус-*

ство), что свидетельствует о диффузности и многомерности данного концепта. Семантическое пространство структурировано через несколько взаимосвязанных, но дифференцированных доменов: эстетико-оценочного (*гений, талант, заслуга, критик*), профессионально-рефлексивного (*ремесло, сложность, очарование*), экзистенциального (*жизнь, мечта*), метапоэтического (*оценивать, пародия, бестолковый*).



Илл. 6. Квази-синонимы для слова «поэзия» в канонической и наивной поэзии

Наивная поэзия, напротив, характеризуется более высокой концентрацией связей и организуется преимущественно вокруг формально-институциональных параметров: формальных аспектов и характеристик (*стихотворный, стихотворение, рифма, содержание, стиль, проза*), литературных институций (*литературный, литература, автор, писатель*). Структурная дифференциация подтверждает наш более общий тезис о принципиально различных моделях концептуализации поэзии в двух традициях: каноническая поэзия оперирует размытыми, многомерными семантическими полями с акцентом на эстетическую и философскую рефлексию, тогда как наивная фокусируется на более конкретных, формально-институциональных аспектах литературного производства.

Проведенный анализ демонстрирует систематические различия в семантических структурах наивной и канонической поэзии. Наблюдаемые различия могут быть обусловлены как специфическими для каждой традиции способами концептуализации, так и историко-культурными факторами. Особый интерес представляют случаи, когда при сохранении общего ядра ближайших соседей происходит существенная реорганизация семантического поля, что может свидетельствовать о различных функциональных акцентах в концептуализации одних и тех же лексем.

Системный подход к анализу векторных моделей и сочетание разных аналитических инструментов — от базовых статистических метрик и кластерного анализа до рассмотрения семантических сдвигов отдельных слов — позволяет распознавать структурные закономерности, не всегда доступные при традиционном «медленном чтении». Он позволяет количественно измерить такие интуитивно ощущаемые категории, как «предсказуемость», «литературность» или «метафоричность» текста, через статистически значимые закономерности, а не с помощью примеров из текста, набор которых при значительном объеме сопоставляемых корпусов неизбежно будет субъективным и неполным. Применение такого аналитического подхода для масштабного сопоставления различных литературных традиций, направлений и жанров пока ограничено состоянием существующих корпусов и потребует целенаправленной работы по созданию репрезентативных поэтических собраний, отражающих разнообразие поэтических практик. Очевидно, что объемность предложенному анализу мог бы придать контекст современной русской поэзии за пределами сайта «Стихи.Ру», однако пока такого корпуса в распоряжении исследователей нет.

Библиография / References

Бонч-Осмоловская А.А., Орехов Б.В. Некоторые применения корпусных методов к наивной поэзии // Статьи на случай: Сборник в честь 50-летия Р.Г. Лейбова. М.: Объединенное гуманитарное издательство Ruthenia.ru, 2013. С. 22–36.

(*Bonch-Osmolovskaya A.A., Orekhov B.V.* Nekotorye primeneniya korpusnykh metodov k naivnoy poezii // Stat'i na sluchay: Sbornik v chest' 50-letiya R.G. Leybova. Moscow, 2013. P. 22–26.)

Левченко М.А. Индустриальная свирель: Поэзия Пролеткульта 1917–1921 гг. СПб.: СПГУТД, 2007.

(*Levchenko M.A.* Industrial'naya svirel': Poeziya Proletkulta 1917–1921 gg. Saint Petersburg, 2007.)

Лейбов Р., Орехов Б.В. Между политикой и поэтикой: топика Крыма в современной русскоязычной наивной лирике // Шаги / Steps. 2022. Т. 8. № 2. С. 205–232.

(*Leybov R., Orekhov B.V.* Mezhdu politikoy i poetikoy: topika Kryma v sovremennoy russkoyazychnoy naivnoy lirike // Shagi / Steps. 2022. Vol. 8. № 2. P. 205–232.)

Лурье М.Л. О феномене наивного сочинительства // «Наивная литература»: исследования и тексты / Сост. С.Ю. Неклюдов. М.: Московский общественный научный фонд, 2001. С. 15–28.

(*Lur'e M.L.* O fenomene naivnogo sochinitel'stva // «Naivnaya literatura»: issledovaniya i teksty / Ed. by S.Yu. Neklyudov. Moscow, 2001. P. 15–28.)

Маслинский К.А. Сто лет счастья в детской литературе (1920–2020): сталинский канон и его долгосрочные последствия // Шаги / Steps. 2022. Т. 8. № 4. С. 226–247.

(*Maslinskiy K.A.* Sto let schast'ya v detskoy literature (1920–2020): stalinskiy kanon i ego dolgosrochnye posledstviya // Shagi / Steps. 2022. Vol. 8. № 4. P. 226–247.)

Моретти Ф. Дальнее чтение / Пер. с англ. А. Вдовина, О. Собчука, А. Шели; науч. ред. перевода И. Кушнарева. М.: Издательство Института Гайдара, 2016.

(*Moretti F.* Distant Reading. Moscow, 2016 — In Russ.)

Неклюдов С.Ю. От составителя // «Наивная литература»: исследования и тексты / Сост. С.Ю. Неклюдов. М.: Московский общественный научный фонд, 2001. С. 4–14.

(*Neklyudov S.Yu.* Ot sostavitelya // «Naivnaya literatura»: issledovaniya i teksty / Ed. by S.Yu. Neklyudov. Moscow, 2001. P. 4–14.)

Орехов Б.В. Стихи и проза через призму дистрибутивной семантики // Острова любви БорФеда: сборник к 90-летию Бориса Федоровича Егорова. СПб.: Росток, 2016. С. 652–655.

(*Orekhov B.V.* Stikhi i proza cherez prizmu distributivnoy semantiki // Ostrova lyubvi BorFeda: sbornik k 90-letiyu Borisa Fedorovicha Egorova. Saint Petersburg, 2016. P. 652–655.)

Antoniak M., Mimno D. Evaluating the Stability of Embedding-Based Word Similarities // Transactions of the Association for Computational Linguistics. 2018. Vol. 6. P. 107–119.

Bolukbasi T., Chang K.-W., Zou J.Y., Saligrama V., Kalai A.T. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings // Advances in Neural Information Processing Systems. Barcelona: Curran Associates, 2016. P. 4349–4357.

Bybee J.L. Frequency of Use and the Organization of Language. New York: Oxford University Press, 2007.

Eder M., Šeļa A. One Word to Rule them All: Understanding Word Embeddings for Authorship Attribution // Digital Humanities 2022 Combined Abstracts. Tokyo: Alliance of Digital Humanities Organizations, 2022. P. 199–201.

Eger S., Mehler A. On the Linearity of Semantic Change: Investigating Meaning Variation via Dynamic Graph Models // Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) / Ed. by K. Erk, N.A. Smith. Berlin: Association for Computational Linguistics, 2016. P. 52–58.

Ehrmanntraut A., Hagen T., Konle L., Jannidis F. Type- and Token-Based Word Embeddings in the Digital Humanities // Proceedings of the Computational Humanities Research Conference 2021 (CHR 2021). Amsterdam; Antwerp: CEUR Workshop Proceedings, 2021. P. 16–38.

Ehrmanntraut A., Hagen T., Jannidis F., Konle L., Kröncke M., Winko S. Modeling and Measuring Short Text Similarities. On the Multi-Dimensional Differences between German Poetry of Realism and Modernism // Journal of Computational Literary Studies. 2022. Vol. 1. № 1. P. 1–30.

Firth J.R. A Synopsis of Linguistic Theory, 1930–1955 // Studies in Linguistic Analysis. Oxford: Blackwell, 1957.

Fomin V., Bakshandaeva D., Rodina J., Kutuzov A. Tracing Cultural Diachronic Semantic Shifts in Russian Using Word Embeddings: Test Sets and Baselines // arXiv:1905.06837. 2019 (URL: https://arxiv.org/abs/1905.06837).

Garg N., Schiebinger L., Jurafsky D., Zou J. Word Embeddings Quantify 100 Years of Gender and Ethnic Stereotypes // Proceedings of the National Academy of Sciences. 2018. Vol. 115. № 16. P. E3635–E3644.

Grayson S., Mulvany M., Wade K., Meaney G., Greene D. Exploring the Role of Gender in 19th Century Fiction Through the Lens of Word Embeddings // Language, Data, and Knowledge / Ed. by J. Gracia, F. Bond, J.P. McCrae, P. Buitelaar, C. Chiarcos, S. Hellmann. Cham: Springer, 2017. P. 358–364.

Hamilton W.L., Leskovec J., Jurafsky D. Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change // Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) / Ed. by K. Erk, N.A. Smith. Berlin: Association for Computational Linguistics, 2016. P. 1489–1501.

Hatzel H.O., Stiemer H., Biemann C., Gius E. Machine Learning in Computational Literary Studies // Information Technology. 2023. Vol. 65. № 4–5. P. 200–217.

Hengchen S., Ros R., Marjanen J. A Data-Driven Approach to the Changing Vocabulary of the «Nation» in English, Dutch, Swedish and Finnish Newspapers, 1750–1950 // Book of Abstracts of DH2019. Utrecht: Utrecht University, 2019. P. 125–130.

Hengchen S., Ros R., Marjanen J., Tolonen M. A Data-Driven Approach to Studying Changing Vocabularies in Historical Newspaper Collections // Digital Scholarship in the Humanities. 2021. Vol. 36. Issue Supplement 2. P. ii109–ii126.

Kulkarni V., Al-Rfou R., Perozzi B., Skiena S. Statistically Significant Detection of Linguistic Change // Proceedings of the 24th International World Wide Web Conference / Ed. by A. Gangemi, S. Leonardi, A. Panconesi. Florence: ACM, 2015. P. 625–635.

Kutuzov A., Øvrelid L., Szymanski T., Velldal E. Diachronic Word Embeddings and Semantic Shifts: a Survey // Proceedings of the 27th International Conference on Computational Linguistics / Ed. by E.M. Bender, L. Derczynski, P. Isabelle. Santa Fe: Association for Computational Linguistics, 2018. P. 1384–1397.

Mikolov T. et al. Efficient Estimation of Word Representations in Vector Space // arXiv:1301.3781. 2013 (URL: https://arxiv.org/abs/1301.3781).

Mikolov T., Sutskever I., Chen K., Corrado G., Dean J. Distributed Representations of Words and Phrases and their Compositionality // Proceedings of the 27th International Conference on Neural Information Processing Systems. Vol. 2 / Ed. by C.J.C. Burges et al. Red Hook: Curran Associates, 2013. P. 3111–3119.

Pagel M., Atkinson Q.D., Meade A. Frequency of Word-Use Predicts Rates of Lexical Evolution throughout Indo-European History // Nature. 2007. Vol. 449. № 7163. P. 717–720.

Pennington J., Socher R., Manning C. GloVe: Global Vectors for Word Representation // Proceedings of the 2014 Conference on Empiri-

cal Methods in Natural Language Processing (EMNLP) / Ed. by A. Moschitti, B. Pang, W. Daelemans. Doha: Association for Computational Linguistics, 2014. P. 1532–1543.

Perri V., Qarkaxhija L., Zehe A., Hotho A., Scholtes I. One Graph to Rule them All: Using NLP and Graph Neural Networks to Analyse Tolkien's Legendarium // Proceedings of the Computational Humanities Research Conference 2022. Vol. 3290. Antwerp: CEUR Workshop Proceedings, 2022. P. 291–317.

Ros R. Conceptual Vocabularies and Changing Meanings of «Foreign» in Dutch Foreign News (1815–1914) // Book of Abstracts of DH2019. Utrecht: Utrecht University, 2019. P. 230–238.

Ros R., van Eijnatten J. Disentangling a Trinity: A Digital Approach to Modernity, Civilization and Europe in Dutch Newspapers (1840–1990) // Book of Abstracts of DH2019. Utrecht: Utrecht University, 2019. P. 42–47.

Schönemann P.H. A Generalized Solution of the Orthogonal Procrustes Problem // Psychometrika. 1966. Vol. 31. P. 1–10.

Shin H. Analyzing the Positive Sentiment towards the Term «Queer» in Virginia Woolf through a Computational Approach and Close Reading // Journal of Computational Literary Studies. 2022. Vol. 1. № 1. P. 1–26.

Stoltz D.S., Taylor M.A. Cultural Cartography with Word Embeddings // Poetics. 2021. Vol. 88. P. 1–14.

Tripto N.I., Ali M.E. The Word2vec Graph Model for Author Attribution and Genre Detection in Literary Analysis // arXiv:2310.16972. 2023 (URL: https://arxiv.org/abs/2310.16972).

Turney P.D., Pantel P. From Frequency to Meaning: Vector Space Models of Semantics // Journal of Artificial Intelligence Research. 2010. Vol. 37. № 1. P. 141–188.

Wendlandt L., Kummerfeld J.K., Mihalcea R. Factors Influencing the Surprising Instability of Word Embeddings // Proceedings of NAACL-HLT / Ed. by M.A. Walker, H. Ji, A. Stent. New Orleans: Association for Computational Linguistics, 2018. P. 2092–2102.

Wevers M. Using Word Embeddings to Examine Gender Bias in Dutch Newspapers, 1950–1990 // arXiv:1907.08922. 2019 (URL: https://arxiv.org/abs/1907.08922).