

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ
РОССИЙСКОЙ ФЕДЕРАЦИИ
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ
УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«АЛТАЙСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»

На правах рукописи



Остапенко Светлана Валерьевна

МОДЕЛИРОВАНИЕ СОДЕРЖАНИЯ СООБЩЕНИЯ В ПРОЦЕССЕ
АВТОМАТИЧЕСКОЙ ГЕНЕРАЦИИ ТЕКСТА: ЛИНГВО-ИНФОРМАЦИОННЫЙ
АСПЕКТ

Специальность 5.9.8. – Теоретическая, прикладная и сравнительно-
сопоставительная лингвистика (филологические науки)

Диссертация на соискание учёной степени
кандидата филологических наук

Научный руководитель:
доктор филологических наук,
профессор
Халина Наталья Васильевна

Барнаул – 2025

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ	4
ГЛАВА 1. ПРОБЛЕМЫ ТЕКСТОВОЙ ГЕНЕРАЦИИ	14
1.1. Деривационные и генерационные процессы текстообразования	14
1.2. От цифрового текста к сгенерированному тексту.....	19
1.3. Феномен сгенерированного текста как объект лингвистического исследования.....	22
1.4. Оптимизация и генерация как механизмы функционирования текстов в цифровой среде.....	28
1.5. Методология анализа сгенерированного текста и оценка его качества.....	33
Выводы по первой главе	47
ГЛАВА 2. СТРАТЕГИИ ИНФОРМАЦИОННОГО ТЕКСТОПОРОЖДЕНИЯ: КОММУНИКАЦИЯ С ДИАЛОГОВЫМИ АГЕНТАМИ.....	49
2.1. Компьютерно-опосредованная коммуникация и диалоговая лингвистика ...	49
2.2. Интерфейс автогенерации текстов: лингвистические аспекты анализа промтов.....	55
2.3. Промт-инжиниринг как стратегия текстопорождения	64
2.4. Информационная модель содержания сообщения в контексте генерации текста	73
2.5. Лингвокогнитивные аспекты восприятия и понимания сгенерированных текстов	84
Выводы по второй главе	95
ГЛАВА 3. ЛИНГВО-ИНФОРМАЦИОННОЕ МОДЕЛИРОВАНИЕ СГЕНЕРИРОВАННОГО ТЕКСТА	99
3.1. Лингво-информационная модель коммуникации в системе «человек – диалоговый агент».....	99
3.2. Тематическое моделирование сгенерированных диалоговых текстов	109
3.3. Структурно-композиционные особенности сгенерированных диалоговых текстов	119
3.4. Морфолого-синтаксическая организация сгенерированных диалогов.....	128
3.5. Сравнительный анализ сгенерированных и естественных диалогов	133
3.6. Методика атрибуции сгенерированного диалогового текста на основе лингво-информационной модели	148

Выводы по третьей главе.....	158
ЗАКЛЮЧЕНИЕ	160
СПИСОК ЛИТЕРАТУРЫ	167
Приложение 1.....	191
Приложение 2.....	209
Приложение 3.....	229
Приложение 4.....	247

ВВЕДЕНИЕ

В условиях активного развития больших языковых моделей и технологий генерации контента с использованием нейросетей и их экспансией в разных сферах публичной коммуникации остро встает вопрос о лингвистическом статусе и специфике репрезентации информационного содержания в языковых продуктах генерации – искусственных (сгенерированных) текстах на естественном языке.

Искусственный текст, лишенный лингвопрагматических признаков, характерных для антропологического текста, одновременно выступает как феномен автоматического текстопорождения, не связанный с речевой деятельностью субъекта. Стохастические модели порождения речи и теория генеративной грамматики, составляющие лингвистический базис прикладных задач текстогенерации, рассматривают продукт генерации в формальном аспекте, редуцируя или выводя за пределы научной рефлексии вопросы семантики и прагматики, значимые для функционирования сгенерированных текстов в различных типах дискурсов – от коммерческого до научного.

В связи с этим настоящее исследование направлено на изучение сгенерированных текстов как продуктов нового типа коммуникации в системе «человек – машина», опосредованных развитием генеративных алгоритмов посредством создания их лингво-информационной модели, включающей как формально-структурные, так и семантико-прагматические аспекты.

Актуальность работы, таким образом, обусловлена противоречием между экспансией сгенерированного текста в различных сферах коммуникации и недостаточной изученностью его коммуникативно-речевой природы. Разработка лингво-информационной модели сгенерированного текста является значимой в теоретическом и прикладном аспектах по ряду причин: во-первых, ввиду отсутствия комплексной методики лингвистического описания и анализа продуктов текстогенерации; во-вторых, в связи с недостаточностью теоретической разработанностью проблематики искусственного (сгенерированного) текста в современной лингвистике. Помимо указанных факторов, актуальность работы

также обусловлена использованием методов автоматической обработки и анализа естественного языка, релевантных цифровой природе изучаемого феномена – текста, порожденного на основе алгоритмических процессов.

Степень разработанности темы. Теоретическую основу исследования составили научные труды отечественных и зарубежных ученых в области генеративной лингвистики – N. Chomsky, R. Jackendoff, H. Lasnik, M. Saito, J.-Y. Pollock, Mark C., Baker L., K. Hale, С. К. Шаумяна; деривационной теории текстообразования – Л. Н. Мурзина, Н. Д. Голева, Ю. В. Трубниковой, Н. Н. Шпильной; диалогической лингвистики в аспекте компьютерно-опосредованной коммуникации – А. А. Барковича, Л. Ю. Щипициной, Н. Н. Шпильной; теории автоматической генерации текста – Л. Н. Иорданской, Е. Г. Соколовой, М. В. Болдасова, Е. И. Большаковой; компьютерной лингвистики и автоматической обработки естественного языка – И. А. Мельчука, И. М. Богуславского, Ю. Д. Апресяна, А. С. D. Manning, H. Schütze, A. Clark, C. Fox, S. Lappin, G. Neubig, J. Eisner, P. Koehn, M. Dredze; корпусной лингвистики – В. А. Плунгяна, Е. Е. Голубковой, А. А. Кибрика, Е.А. Красиной, М. Л. Новиковой, О. Н. Ляшевской, W. N. Francis, H. Kucera; лингвистического информационного моделирования текста и дискурса – В. Я. Яцко, F. S. Eigner; мультимодального текста и дискурса – R. Barthes, M.A.K. Halliday, T. van Leeuwen, G. Kress; человеко-машинной коммуникации – S. Jones, D. J. Gunkel, A. P. Edwards, J. Gibbs, G. L. Kirkwood, C. Fang, J. N. Wilkenfeld, L. Fortunati.

Изучение и анализ источников по тематике исследования подтверждает наличие научно обоснованной теоретико-методологической базы, но в то же время указывает на существование лакун во многих теоретических и прикладных аспектах, связанных с разработкой методологии анализа сгенерированного текста, лингвокогнитивными аспектами его восприятия и интерпретации, а также сопоставлением сгенерированного и антропологического текстов для выявления надежных параметров атрибуции лингвистических признаков генерации.

Методологическая основа исследования носит комплексный характер, базируясь на сочетании принятых в современной лингвистике подходов и

методов. Единство формально-структурного и семантического подходов дополняется элементами лингвокоммуникативной и дескриптивной парадигм анализа, обеспечивая всестороннее изучение рассматриваемого феномена. В диссертационном исследовании использованы следующие общенаучные и частнонаучные методы познания: описательный метод (включает анализ и систематизацию имеющихся теоретических и эмпирических данных по теме исследования); метод лингвистического эксперимента (в аспекте исследования влияния стратегий текстопорождения и приемов промт-инжиниринга на формально-языковые характеристики генерируемых текстов); лингвистическое моделирование (в части разработки лингво-информационной модели сгенерированного текста); метод корпусных исследований (при анализе выборок, включающих большие массивы текстовых данных); методы автоматической обработки естественного языка с использованием языка программирования Python (для автоматизации процессов количественного и качественного исследования корпусов текстов); коммуникативно-речевой и формальный анализ текстов (в ходе выявления лингвистических параметров сгенерированных текстов, а также при последующем сопоставительном анализе сгенерированных и антропологических текстов), метод лингвокоммуникативного анализа (при исследовании коммуникативных целей, влияющих на параметры запросов для генерации текстов); метод лингвокогнитивного анализа (в части выявления механизмов репрезентации смысла в сгенерированных текстах); метод сравнительно-сопоставительного анализа (для выявления специфики сгенерированного и антропологического текстов в аспекте анализа их структурно-композиционных, языковых и тематических особенностей); метод системного описания (с целью обобщения выводов по теме исследования и представления полученных результатов).

Объектом исследования является сгенерированный текст, полученный в результате коммуникативного взаимодействия субъекта (пользователя) и диалоговых систем на основе генеративных алгоритмов (моделей). **Предметом** исследования выступают параметрические характеристики, извлеченные из

сгенерированного текста и обобщенные в виде лингвистической модели, воспроизводящей их структурно-семантические и коммуникативно-речевые особенности.

В соответствии с **гипотезой**, которая была выдвинута в данном исследовании, лингво-информационная модель сгенерированного текста, включающая его тематические, структурно-композиционные и морфолого-синтаксические параметры (паттерны), может быть использована в качестве основы для разработки лингвистической методики атрибуции сгенерированных текстов без обращения к формальным (стилометрическим, статистическим, машинно-алгоритмическим) подходам.

Цель исследования – выявить и проанализировать лингвистические характеристики (параметры) сгенерированного текста, полученного в результате взаимодействия пользователя с системами языковой генерации на основе больших языковых моделей, для разработки обобщенной лингво-информационной модели его содержания, отражающей особенности и закономерности данного типа текстов.

В соответствии с выбранной целью были сформулированы следующие задачи исследования:

1. Выявить лингвистические механизмы, лежащие в основе генерационных процессов, путем изучения их генезиса в аспекте трансформационной грамматики и деривационной теории текстообразования.

2. Определить феномен сгенерированного (искусственного) текста как объект лингвистического исследования, его отличия от цифрового текста и специфику методов его лингвистического анализа в методологическом поле современной лингвистики;

3. Выявить и проанализировать стратегии текстопорождения в коммуникации с диалоговыми системами взаимодействия в аспекте промт-инжиниринга и проблемы оценки коммуникативного качества сгенерированного текста;

4. Исследовать принципы, лежащие в основе информационной модели содержания сообщения в контексте генерации текста и определяющие особенности его восприятия и понимания;

5. Проанализировать сгенерированные тексты для выявления их характеристик и разработки лингво-информационной модели текста как продукта взаимодействия человека и машины, основанного на современных больших языковых моделях и генеративных алгоритмах;

6. Сопоставить паттерны лингво-информационной модели сгенерированного текста с релевантными примерами естественной диалогической коммуникации для выявления специфики его природы;

7. Предложить методику атрибуции сгенерированных текстов на основе выявленных параметров без использования автоматических методов атрибутирования.

Научная новизна диссертационной работы заключается в том, что в исследовании впервые предложена лингво-информационная модель сгенерированного текста, отличающаяся комплексным подходом к продукту автоматического текстопорождения на основе деривационной теории текстообразования, а также использовано сочетание количественных и качественных аспектов анализа, что позволяет разработать методику атрибуции сгенерированного текста с опорой на лингвистические параметры, выявляемые в его формально-содержательной структуре.

Теоретическая значимость исследования определяется следующими аспектами: установлен генезис феномена автоматической генерации текста в концепции трансформационной грамматики и деривационной теории текстообразования, представляющих в качестве методологической основы изучения процессов текстогенерации в лингвистическом аспекте; рассмотрена специфика сгенерированного текста с точки зрения механизмов оптимизации текстовых структур и трансформации стратегий текстопорождения, что позволяет определить лингвистическую специфику промт-инжиниринга как процесса лингвокогнитивного моделирования информационного содержания генерируемого

текста, выступающего продуктом коммуникативного взаимодействия в системе «человек – диалоговый агент на основе генеративных моделей»; предложена лингво-информационная модель описания и анализа сгенерированного текста на основе сравнения её характерных параметров с естественными диалогическими текстами, что вносит вклад в развитие теоретических подходов к исследованию лингвистических механизмов автоматической текстогенерации; разработана методика атрибуции сгенерированного текста, позволяющая определять признаки генерации текстового содержания на основе квалификативного подхода.

Практическая значимость диссертационного исследования состоит в том, что его результаты позволяют внедрить предложенный подходы к атрибуции признаков генерации в тексте в существующую методологию атрибутирования, повышая достоверность и обоснованность результатов проверки текстового содержания на предмет признаков генерации с использованием генеративных алгоритмов; использовать лингво-информационную модель сгенерированного текста для прогнозирования его содержания в процессе промт-инжиниринга для улучшения качества генерируемого контента; использовать предложенную модель описания и атрибуции для исследования больших массивов сгенерированных текстов специалистами в области компьютерной и корпусной лингвистики, а также в дидактических целях при обучении студентов в сфере лингвистики, медиатехнологий и компьютерной лингвистики принципам анализа текстового контента генерационной природы; применять предложенную методику сравнительно-сопоставительного анализа сгенерированного и естественного диалогических текстов для проведения сравнительных исследований сгенерированных текстов и естественных речевых произведений разных жанров и стилей.

Материалом исследования послужили русскоязычные сгенерированные тексты, отобранные в соответствии с механизмами их производства, а именно: текстогенерации с помощью диалоговых агентов – интерфейсов, способных к созданию текста на основе запроса пользователя (субъекта или другого алгоритма). Из источника – публичного датасета со сгенерированными текстами

для коммерческого и научного использования общим объемом 448506 диалогов, включающих единство текста-стимула (промта) и текста-реакции (продукта генерации), методом случайной выборки отбирались тексты по критерию коммуникативно-тематической направленности (персонализированные диалоги). Для решения задачи сравнительного анализа сгенерированных и естественных текстов были использованы русскоязычные диалогические тексты устной непубличной коммуникации, размещенные в Национальном корпусе русского языка и агрегированные по следующим параметрам: устная непубличная речь, атрибут авторства – смешанный гендерный признак (женский и мужской). Объем подкорпуса составляет 1940 диалогических текстов, включающих не менее 10 реплик каждый.

Единицами анализа выступают: 1) корпус из 2000 сгенерированных диалогов персонализированного (непубличного) типа коммуникации из указанного датасета, принадлежащих разным нейросетевым маскам персонажей и включающим от 6 до 20 реплик; 2) корпус из 1940 диалогов естественной непубличной коммуникации (представлены в устном подкорпусе Национального корпуса русского языка). В ходе исследования была оценена пригодность русскоязычных диалогических текстов устной непубличной коммуникации для сопоставления со сгенерированными текстами. Естественные тексты признаны полностью пригодными для исследования по формально-структурным, тематическим и грамматическим параметрам, что обеспечивает достаточный уровень сопоставимости со сгенерированными диалогами. В отношении акцентологических, просодических и мелодических характеристик естественные тексты признаны ограниченно пригодными, так как эти особенности устной речи не могут быть адекватно сопоставлены с особенностями сгенерированных текстов, имитирующих естественную письменную речь. Поэтому анализ сосредоточен на формальных, структурных и тематических параметрах, исключая специфические особенности устных диалогов. Два корпуса подвергнуты сопоставительному анализу по разработанной методике для выявления

специфических параметров лингво-информационной модели сгенерированных текстов.

На защиту выносятся следующие **положения**:

1. Генерация текста с использованием нейросетей интегрирует принципы трансформации и динамического преобразования текстовых структур, заложенные в теоретические модели деривации текста, на качественно новом техническом уровне.
2. Сгенерированный текст представляет собой лингвистический объект нового типа, в котором антропологические стратегии текстопорождения заменяются алгоритмическими, что наряду с автоматизацией создания и размножением текстов типовой структуры, параметризацией и прогнозированием содержания приводит к отчуждению текста от речевой деятельности субъекта и утрате индивидуально-авторских черт.
3. Промт-инжиниринг как методология разработки и оптимизации текстовых запросов для взаимодействия с нейросетевыми алгоритмами улучшает коммуникативное качество продуктов текстогенерации и трансформирует механизмы организации информации и восприятия сгенерированных текстовых сообщений. Когнитивная модель интерпретации сгенерированного текста задается выбранной стратегией промт-инжиниринга в соответствии с его жанровым аспектом и концептуализируется с помощью пропозициональной и семантической структуры генерируемого сообщения.
4. Комплексная лингво-информационная модель сгенерированного текста включает тематические, структурно-композиционные и морфолого-синтаксические компоненты, позволяющие выявить специфические характеристики автоматического текстопроизводства, включая паттерны речевых актов, маркеры переходов между репликами и лексико-грамматическую организацию диалога. Сравнение сгенерированных и естественных диалогов по параметрам тематического разнообразия, частотности субстантивной лексики, значения индекса дейктичности,

наличия маркеров речевой экспрессии, уровня формально-грамматической структурированности подчеркивает качественные различия между двумя типами текстов, выявляя характерные паттерны для их идентификации.

5. Методика атрибуции сгенерированного диалогового текста основана на комплексном анализе, позволяющем выявлять признаки генерации. Процесс атрибуции включает оценку количественных и качественных характеристик текстов и позволяет получать надежные результаты при атрибуции сгенерированного контента и (или) наличия его признаков в тексте, дополняющие данные программных алгоритмов и сервисов детекции сгенерированного контента.

Апробация работы. Основные положения диссертационного исследования представлены в 9 научных статьях, 3 публикации размещены в рецензируемых научных изданиях, рекомендованных Высшей аттестационной комиссией. Промежуточные результаты исследовательской работы представлены в докладах на научных конференциях: IV Транснациональной конференции «Алтайский медиабарометр» (30–31 октября 2024 г., г. Барнаул), Международной научной конференции «Русская литература в России и в мире-3», посвященной 225-летию со дня рождения А. С. Пушкина (17-19 октября 2024 г., г. Улан-Удэ), Международной научно-практической конференции «Индустрия 4.0» (23-29 марта 2025 г., г. Сочи), IV Международной научно-практической конференции «Онтологическая безопасность азиатско-алтайского фронта: актуальные проблемы транснациональных медиаисследований» (9-11 апреля 2025 г., г. Барнаул), VII Международной научно-практической конференции «Актуальные проблемы языкознания и методики преподавания иностранных языков» (8-9 апреля 2025 г., г. Челябинск), III Международной научной конференции «Текст – дискурс – культура» (24-26 апреля 2025 г., г. Краснодар), XIV Международной научно-практической конференции «Цифровое общество: научные инициативы и новые вызовы» (16 июня 2025 г., г. Москва).

Структура работы. Диссертация состоит из введения, трех глав, заключения, библиографического списка (включающего 227 научных источников) и 4 приложений.

ГЛАВА 1. ПРОБЛЕМЫ ТЕКСТОВОЙ ГЕНЕРАЦИИ

1.1. Деривационные и генерационные процессы текстообразования

Проблематика текстообразования и динамический аспект изучения текста занимают центральное место в современном языкознании. Исследования в этой области направлены на понимание механизмов, формирующих тексты различной природы. Одним из ключевых направлений изучения текстообразования в динамическом аспекте является теория деривационного моделирования, которая исторически восходит к идеям В. фон Гумбольдта и А. А. Потебни о ничем не ограниченной творческой природе порождения речи на базе существующего языкового материала. Современные исследования в этой области связаны с теорией порождающей грамматики Н. Хомского (N. Chomsky) [Chomsky 1957, 1970], а также с господствующим в лингвистике со второй половины XX века функциональным (процессуальным) подходом к изучению текста.

В лингвистике возникновение термина «деривация» связывают с исследованиями Е. Куриловича в 1930-х-1940-х гг., который использует его для описания словообразовательных процессов [Курилович 2000]. Этот этап развития теории деривации можно считать начальным, трактующим деривацию достаточно узко, применительно к морфологическому уровню языковой системы. Различая два типа деривации – лексическую и синтаксическую, Е. Курилович приходит к выводу, что изменение значения языковой единицы (уровень семантики) не всегда влечет за собой трансформацию его формы и наоборот: изменение формы (синтаксической структуры) необязательно влечет за собой изменение смысловой структуры предложения. Эти идеи оказались продуктивными для развития функционального подхода к деривации, которая стала рассматриваться как процесс изменения функции элемента в структуре текста.

Важной фигурой в развитии теории деривации стал американский лингвист Ноам Хомский, создатель концепции трансформационной грамматики. Н. Хомский предложил рассматривать синтаксическую деривацию как серию трансформаций, которые преобразуют глубинные структуры предложения в

поверхностные. Глубинные структуры, как их понимает Н. Хомский, являются лингвокогнитивными схемами, языковыми «шаблонами», общими для множества предложений и связанными когнитивным механизмом усвоения языка (language acquisition device), тогда как поверхностные структуры выступают в качестве конкретных реализаций этих шаблонов в виде грамматически оформленных предложений [Chomsky 1957, 2001]. Трансформации, согласно Хомскому, являются базовыми операциями по преобразованию глубинных структур в поверхностные и включают в перемещение, добавление, удаление и перестановки элементов синтаксической структуры, которые, по мысли американского лингвиста, объясняют не только разнообразие синтаксических вариантов, но и возможность образование новых, ранее не встречавшихся в речевом опыте субъекта конструкций.

Последовательное развитие идей генеративизма на отечественной почве находим в трудах И. А. Мельчука [Мельчук 1999] применительно к разработанной им концепции «Смысл – Текст» (СТ). Согласно разработанной И. А. Мельчуком теории, язык рассматривается как конечная система правил, обеспечивающих соответствие между множествами смыслов (семантических представлений) и множествами текстов (фонетических представлений). Это соответствие характеризуется многозначностью и дискретностью отношений между элементами и позволяет языку функционировать как сложным образом организованная система коммуникации. Методология теории СТ восходит к идеям трансформационной грамматики Н. Хомского и базируется на глубинно-поверхностных правилах (трансформациях). Другим важным аспектом методологии теории СТ является морфологическая декомпозиция, предусматривающая разложение слов на составляющие морфемы с учетом возможных алломорфных вариаций и обеспечивающая возможность грамматической трансформации на уровне языковой единицы при изменении контекста её употребления.

Теория СТ вкупе с трансформационной грамматикой Н. Хомского была положена в основу лингвистических механизмов алгоритмов машинного перевода

и генерации текста. Основные принципы генерации текста компьютерными алгоритмами, восходящие к соответствующим концепциям теории СТ, базируются на многовариантном преобразовании смысла в текст посредством ограниченного набора формальных правил их сочетания, что соответствует основной идее теории СТ.

Трансформационные модели текстообразования, предложенные генеративной лингвистикой, были пересмотрены в парадигме постгенеративизма, которая характеризуется подходом к тексту как самоорганизующейся системе, функционирующей и развивающейся в рамках синергетических принципов. В отечественном языкознании этот переход связан с развитием парадигмы дериватологии и её экспансией на уровень производства текста. Поворот к деривационной модели текстообразования намечается в исследованиях пермской лингвистической школы начиная с 1970-х годов в трудах Л. Н. Мурзина [Мурзин 1974] и его учеников (С. Ю. Адливанкин, В. А. Мишланов, И. Ю. Черепанова, Л. М. Алексеева), которые делают акцент на динамических аспектах языка и рассматривают его как систему, обладающую потенциально бесконечными возможностями для развития. Основопологающим понятием теории деривации в пермской школе дериватологии является «мысле-речевое действие», описывающее процесс порождения предикативных единиц на основе семантико-синтаксических принципов, заложенных в языковых моделях [Мурзин 1982]. Деривационные процессы предполагают последовательное развертывание этих принципов и позволяют создавать аутентичные высказывания на естественном языке.

В русле теории дериватологии развиваются два аспекта изучения моделей деривации применительно к текстовым структурам – синтаксическая и лексическая. Основы синтаксической деривации заложены в трудах Л. Н. Мурзина [Мурзин 1974]. В лексико-деривационной концепции текста Ю. В. Трубниковой [Трубникова 2008] фокус с синтаксиса смещается на лексический уровень, и текст рассматривается как сеть лексико-деривационных рядов, образующих уникальные смысловые конфигурации.

Таким образом, в конце XX – начале XXI вв. в отечественной лингвистике формируется направление, которое можно назвать дериватологическим описанием текста [Голев 1998], [Разина 2012], [Шпильная 2018]. Деривационная модель текстообразования, предложенная Н. Д. Голевым, рассматривает процесс создания текста как преобразование (трансформацию) инварианта (исходного материала) через последовательность производных действий [Голев 1989, 1998]. Процесс преобразования инварианта включает мотивацию и деривацию, определяющие характер вариаций в тексте. Мотивация относится к семантическим связям элементов текста, в то время как деривация включает грамматические и стилистические преобразования, охватывающие уровень текстовой структуры. Модель учитывает динамические аспекты лексической мотивации, подчеркивая, что текст создается через взаимодействие различных уровней языковой системы. Распространяя предложенную модель на вторичные тексты (изложения, переводы и пародии), исследователи рассматривают их как продукт деривационных процессов, основанных на взаимодействии оригинального текста и новых условий его воспроизведения [Голев, Сайкова 2001].

Модели текстообразования, выявленные в рамках исследования процессов деривации текста, репрезентированы в базовых технологиях генерации сообщений, которые вкупе с собственно техническими принципами работы больших языковых моделей (сложная архитектура, глубокое обучение, механизм внимания (attention) для выделения релевантных частей текстовых данных, техники переноса знаний (transfer learning) и др.) позволяют создавать целостные в структурно-семантическом отношении тексты, аутентичные для естественной коммуникации. В Таблице 1 представлены ключевые аспекты деривационной теории текстообразования в сравнении с современными методами автоматической генерации текста, выявляющие основные принципы теории деривации, нашедшие отражение в создании искусственных (сгенерированных) текстов.

Таблица 1. Отражение принципов деривационной концепции текстопорождения в процессах автоматической генерации текста

Основные положения деривационной модели текстообразования	Отражение принципов теории деривации текста в современных моделях текстогенерации
Выделение глубинных и поверхностных структур с акцентом на формально-смысловые аспекты трансформации смыслового инварианта	Использование механизма внимания (attention mechanisms) для определения иерархий зависимостей между словами
Морфологическое разложение (декомпозиция) текста	Модели токенизации и лемматизации, которые представляют текст как последовательность токенов и лемм – минимальных единиц автоматического анализа текстовой структуры
Синтаксическая деривация в развертывании текста	Акцент на синтаксических связях в тексте, реализованных в архитектуре современных моделей- трансформеров
Лексическая деривация в текстопорождении	Моделирование вероятностей появления слов на основе запоминания и анализа предшествующего контекста
Принцип многоуровневости анализа, восприятия и оценки динамической текстовой структуры	Многослойные архитектуры нейросетевых алгоритмов, учитывающие различные паттерны при моделировании текста
Динамическое (процессуальное) представление текста	Рекуррентные нейронные сети и архитектура LSTM, способные обрабатывать длинные текстовые последовательности (выход за пределы слова и предложения и переход к анализу их динамической реализации в контексте)

Генерация текста с использованием современных языковых моделей, основанная на принципах трансформации текстовых структур (паттернов), извлечённых из обучающих текстовых выборок и запросов пользователя, выступает техническим аспектом реализации идей деривации. При этом благодаря современным технологиям эти принципы реализуются на новом уровне с

привлечением больших данных и методов глубокого обучения, отражая эволюцию деривационной модели текстообразования в эпоху погружения языка и речи в цифровой контекст.

1.2. От цифрового текста к сгенерированному тексту

Современная коммуникация активно перемещается в цифровое пространство, формируя дискурсы социальных медиа, интернет-ресурсов, интеллектуальных систем генерации текстов и изображений [Потапова, Курьянова 2021: 6–7]. Возникающие в ходе цифровой коммуникации речевые произведения предстают перед нами в знаково-символической форме, сочетающей семиотические компоненты различной природы: вербальные, визуальные, аудиальные, кинестетические. В междисциплинарных исследованиях, медиалингвистике, дискурсологии такие речевые произведения по-прежнему именуют «текстом», однако подчеркивают их отличия от традиционных текстов, акцентируя внимание на интерактивности, гипертекстуальности, мультимедийности, адаптивности и других свойствах [Рябцева 2017: 390].

Для объективации нового понятийного содержания используются различные термины, поэтому первоначально необходимо определить термин «цифровой текст» в русле современных лингвистических и междисциплинарных исследований. В частности, целесообразно, на наш взгляд, во-первых, разграничить его содержание с функционирующими в научной литературе терминами «креолизованный текст», «поликодовый текст», «медiateкст», «мультимодальный текст», во-вторых, определить его специфику исходя из структурно-семиотических и функционально-прагматических признаков, в-третьих, обозначить специфику цифрового текста как объекта компьютерной лингвистики.

Обратимся к специфике термина «цифровой текст» в сравнении с близкими терминологическими сочетаниями: *поликодовый текст*, *креолизованный текст*, *мультимодальный текст*.

Общепринятое определение креолизованного текста акцентирует внимание на разнородности его «фактуры», которая состоит из вербальной (языковой/речевой) и невербальной (принадлежащей к другим знаковым системам, нежели естественный язык) составляющих [Сорокин, Тарасов 1990]. К креолизованным текстам можно отнести рекламные тексты, комиксы, плакаты, афиши, мемы.

Обладая негомогенной семиотической природой, креолизованный текст может, во-первых, функционировать вне мультимедийной коммуникации, во-вторых, выступать как мономодальный (это касается, например, полностью креолизованных текстов, в которых присутствует только иконическая составляющая), в-третьих, структурно представлять собой дикодовое образование, сочетающее две гетерогенные системы знаков – визуальную и вербальную. Это отличает его от поликодового текста, который изначально характеризуется наличием нескольких негомогенных знаковых систем, функционирующих в составе единого текстового пространства: вербальных (фонетических или графематических), паравербальных (просодических и кинетических), иконических (статистических и динамических), аудиальных, и проч. [Креолизованный текст 2020: 9-15].

Возникновение термина «поликодовый текст» относится к середине 1970-х годов и связано с типологией текстов, которую предложили Г. В. Ейгер и В. Л. Юхт [Ейгер, Юхт 1974], на основании разграничения в структуре текста гетерогенных семиотических компонентов (кодов): кода естественного языка и иной семиотической системы. То есть поликодовость как свойство смешанного текста основывается на разнородности его семиотической структуры, в связи с чем наиболее коммуникативно благоприятная среда функционирования такого текста – мультимедийная коммуникация, представляющая возможности для органичного сочетания различных семиотических систем благодаря техническим механизмам – гиперссылкам, кликабельным изображениям, наличием вставок аудио- и видеофрагментов. В зарубежной лингвистике вместо термина «поликодовый текст» в аналогичном значении употребляется термин

«мультиmodalный текст» [Adami, Kress 2014; Kress 2010; Jewitt 2009; T. van Leeuwen 2005].

Термин «цифровой текст», используемый наряду с понятиями «мультиmodalный текст» и «полюкодový текст», распространился в последние десятилетия в связи с возникновением интереса к цифровой гуманитаристике в целом и цифровой филологии в частности [Чёрнёнская 2019; Северина 2019], однако феномен представления текстовой информации в электронном виде появился задолго до возникновения самого термина. Между тем, как справедливо отмечают исследователи, необходимо различать цифровой текст, который может функционировать только в цифровой среде, от электронного копии обычного текста, который приобретает цифровой формат, но изначально создавался не для нужд цифровой коммуникации [Лебедева, Веселовская, Купрещенко 2020: 80].

В настоящее время цифровой текст рассматривается как феномен дигитализации культуры и атрибутируется через перечисление его характерных особенностей [Лебедева, Веселовская, Купрещенко 2020; Пищальникова 2021]:

создание, распространение и восприятие в цифровой среде;

технологическая опосредованность, высокая тиражируемость;

интерактивность и мультимедийность, предполагающие не только разнообразие способов и каналов представления информации, но и возможности активного взаимодействия с текстом (например, инструменты автоматического поиска, комментирования, редактирования и проч.);

полюкодóвость, или семиотическая гетерогенность фактуры;

мультиmodalность, связанная с позицией реципиента, который при восприятии цифрового текста задействует несколько каналов восприятия.

В качестве общепотребительного аналога понятия «цифровой текст» также используют сочетание «цифровой контент», акцентируя содержательное наполнение различных форм интернет-дискурса.

При этом нельзя согласиться с определением цифрового текста как специфической информационно-коммуникативной единицы, построенной на полюкодóвой основе для репрезентации доминантного авторского смысла

средствами различных семиотических систем (В. А. Пищальникова). Методологически более ценным является представление о цифровом тексте как результате компьютерного кодирования естественной языковой информации, переводе её в машиночитаемый формат [Скоринкин 2016]. Таким образом, понимание и интерпретация цифрового текста возможна в русле современной компьютерной лингвистики, на стыке проблем кодирования текстов на естественном языке, алгоритмизации процессов их обработки, понимания и интерпретации заложенных в них культурных смыслов. Такой подход к пониманию цифрового текста позволяет, на наш взгляд, избежать узкого понимания данного феномена как конструкта, состоящего из переплетения разных семиотических систем, и отразить специфику цифрового текста в русле современных задач прикладной лингвистики, в том числе связанных с компьютерной лингвистикой.

Обработка естественного языка с помощью компьютерных алгоритмов позволяет выделить и формализовать многочисленные параметры, не включающиеся в традиционный лингвистический анализ, способность такой обработки к обнаружению и обобщению скрытых зависимостей выше, чем у человека, хотя и уступает последнему в точности и глубине анализа контекста. В результате текст в компьютерной лингвистике становится не только предметом анализа, но и продуктом генерационных процессов машинных алгоритмов. Генерация текста как новая стратегия текстопорождения в цифровой среде, комбинируясь с традиционными способами создания текстов, выступает как факт коммуникативной практики, влияющий на статус и лингвокоммуникативные характеристики цифрового текста.

1.3. Феномен сгенерированного текста как объект лингвистического исследования

Новая реальность коммуникации в системе «человек – машина» вкупе с развитием больших языков моделей привели к появлению феномена автоматически сгенерированного текста. В настоящее время в научных

исследованиях нет единого общепринятого подхода к номинации такого типа текстов. Их называют «текстами, сгенерированными искусственным интеллектом», «автоматически сгенерированными текстами», «текстами, сгенерированными при помощи нейросети», «искусственными текстами» (А. Л. Айдагулова, Д. В. Белогорская, З. И. Резанова, А. И. Прохоров, К. В. Асадчая, А. О. Исхакова).

Автоматически сгенерированный текст относится к объектам междисциплинарной природы исследования, сочетающим «материю» языка с компьютерными алгоритмами, лежащими в основе механизмов текстопорождения. В связи с совершенствованием нейронных сетей и появлением предварительно обученных моделей-трансформеров, способных генерировать тексты высокого уровня качества (например, GPT, Bert), актуальными становятся вопросы, связанные с атрибуцией сгенерированных текстов, улучшением их качества на основе промт-инжиниринга, а также проверки фактологической точности, этичности и учёта требований контекста. Также в центре внимания современных исследований оказываются вопросы, связанные с изучением коммуникативно-речевой специфики автоматически сгенерированного текста в сравнении с естественными текстами, созданными человеком.

Под автоматически созданным (сгенерированным) текстом (или текстом, созданным с помощью больших языковых моделей, нейросетей, искусственным текстом) следует понимать текст, созданный с помощью специальных алгоритмов или программных генераторов на основе закономерностей естественного языка [Айдагулова 2023; Kolmogorova, Margolina 2024].

Вопрос о разграничении полностью сгенерированных текстов, созданных на основе запроса пользователя, и текстов, полученных в результате автоматического рерайта или обработки (подбора синонимов, перефразирования) исходного текста, остаётся дискуссионным. Очевидно, в первом случае, когда речь идёт о полностью сгенерированных текстах, можно говорить о генерации всех аспектов текстовой структуры – от композиции до содержания и речевого оформления. Во втором случае мы имеем дело с фиксированной композицией и содержанием

(семантикой), которые выражаются с помощью других языковых средств, отличных от исходных.

Сгенерированный текст является объектом изучения в программировании, компьютерной и прикладной лингвистике в различных аспектах. К наиболее значимым относятся следующие направления исследований: атрибуция текста, сгенерированного алгоритмами, использование генерации текстов для создания образовательных, коммерческих, информационных проектов, квантитативный анализ сгенерированного текста, а также выявление лингвистических особенностей продуктов текстовой генерации.

Детекция сгенерированного текста с помощью статистических методов и программно-аппаратных средств предполагает использование математического и статистического аппарата, а также специальных программных комплексов для атрибуции сгенерированного текста по ряду формальных параметров, среди которых А. О. Исхакова выделяет среднее количество знаков пунктуации на 1000 знаков текста; частоту 100 популярных биграмм букв на 1000 символов; частоту служебных слов на 1000 символов; количество уникальных слов на 1000 символов; среднее число слов в предложении; количество грамматических ошибок на 1000 символов; количество предложений на 1000 символов; количество сложноподчиненных предложений на 1000 символов; количество вопросительных предложений на 1000 символов; частоту 100 популярных слов на 1000 символов; частоту 100 популярных 2-грамм слов на 1000 символов; количество слов в семантическом ядре; наличие единства тематики в разных частях текста. Полученные коэффициенты (векторы численных значений формальных характеристик текста) позволяют, по мнению исследователя, выявить инварианты классов сгенерированных и естественных текстов с высокой предсказательной способностью [Исхакова 2016].

Изучение принципов и методики генерации текста на естественном языке с помощью специальных программных комплексов и (или) нейросетей осуществляется также в информационных, образовательных или коммерческих целях. Например, в дидактике алгоритмы генерации применяются для создания

тестовых заданий на основе выборки текстов. Коллектив исследователей [Балашова, Волынская, Макарычев 2016] выделяет следующие подходы к автоматической генерации текста на естественном языке в учебных целях: генерация на основе шаблонов по заданному алгоритму; генерация на основе продукционной модели знаний; генерация на основе семантических сетей; понятийно-тезисная модель, в которой тестовое задание формируется путём выбора контрольной понятийно-тезисной пары и дистракторов на основе других понятий (тезисов). Авторы подчеркивают высокую трудоёмкость формализации знаний с точки зрения обработки текста, который выступает первичным источником знаний, и это затрудняет массовое применение принципов автоматического генерирования учебных заданий на естественном языке.

Выявление специфических формально-количественных характеристик, присущих сгенерированному тексту, на основании его сопоставления с естественными текстами, относится к количественным методам анализа. Так, в исследовании, проведенном Е. В. Тельповым и С. Е. Ларциной, с помощью методов количественной лингвистики были подсчитаны статистические параметры для конкорданса, составленного на материале 80 сгенерированных нейросетью текстов и 80 естественных текстов [Тельпов, Ларцина 2023]. Количественный анализ был произведён на основе метрики для вычисления важности слов в коллекции текстов TF-IDF (Term Frequency – Inverse Document Frequency), коэффициента лексического разнообразия TTR (Type Token Ratio), позволяющего вычислять лексическую сложность текста, а также метрик лексического разнообразия MTL (The Measure of Textual Lexical Diversity) и vocd-D (Vocabulary Diversity). В результате исследования были получены данные, позволяющие говорить о наличии количественных показателей сгенерированного текста в аспекте меньшего лексического разнообразия в сравнении с естественными текстами, а также большей частоты употребления в сгенерированных текстах слов из заголовка, нежели в текстах, созданных человеком.

Анализ лингвистических признаков, присущих сгенерированным текстам, показывает, что к ним могут быть отнесены появление семантических «лакун», заполненных речевыми стереотипами, клише; неоправданные повторы одной и той же мысли в разных абзацах; ослабление логико-смысловых связей между разными частями текста; наличие словообразовательных аномалий, аграмматизмов или фактических ошибок [Айдагулова 2023].

Среди указанных направлений исследования сгенерированного текста для лингвистики интересны те, которые направлены на анализ его лингвистических признаков. Квантитативные методы исследования являются дополнением к качественным, поскольку совершенствование нейросетей и программ-генераторов направлено именно на улучшение количественных показателей (метрик качества моделей), в том числе на повышение индекса лексического разнообразия, уменьшения перплексии, или показателя, отражающего способность модели предсказывать следующее слово в контексте. Поэтому исключительно квантитативные характеристики не могут служить единственным критерием разграничения сгенерированных и естественных текстов.

Представляется продуктивным подход к сгенерированному тексту, основанный на рассмотрении его как одной из разновидностей цифрового текста, функционирующего в интернет-коммуникации. Как разновидность цифрового текста сгенерированный текст обладает следующими характеристиками:

двухединицей системой кодировки, в которой соседствует коммуникативный код естественного языка и машиночитаемый формат представления данных;

возникновением и функционированием в рамках взаимодействия человека и интерфейса машины и (или) программно-аппаратных средств (например, «умный» поиск информации, интернет вещей, голосовые ассистенты и чат-боты);

автоматической генерацией и обусловленными этим фактором высокими темпами производства;

масштабируемостью как способностью к неограниченному производству текстов большого объема на заданные темы;

интерактивностью, понимаемой как релевантность запросу пользователя в ситуации коммуникации человека и компьютерного алгоритма;

мультимодальностью, основанной на сочетании различных типов семиотических ресурсов (вербальный текст в письменной форме в генераторах текстов, вербальный текст в устной форме в голосовых помощниках / ассистентах, невербальные компоненты сообщения – гиперссылки, аудио- и видеоконтент в «умных» системах голосового поиска).

Сгенерированный текст как объект гетерогенной семиотической природы может быть рассмотрен как порождающая модель, вторичная знаковая система, надстраиваемая над языком по определенным правилам образования. Данный подход впервые упомянут в теории трансформационной порождающей грамматики Н. Хомского. В рамках этой теории сформирована концепция приемлемости текста (*linguistic acceptability*), которая противопоставлена традиционной грамматичности (правильности) в аспекте возможности или невозможности того или иного высказывания для порождающей грамматики языка [Chomsky 1970, 2001]. Приемлемость порождающей модели охватывает как семантику, так и грамматику, позволяя отличать предложения, грамматически правильные, но бессмысленные в содержательном аспекте, от высказываний, которые обладают содержательной и формальной приемлемостью. Н. Хомский отмечает, что порождающая грамматика первична по отношению к языковой семантике, которая всегда находится вне формальных аспектов её развертывания. Однако на практике приходится считаться с тем, что потенциальная рекурсивная бесконечность порождающей грамматической модели ограничена именно семантикой, которая не позволяет производить «неприемлемые» в смысловом отношении высказывания. В этом аспекте феномен генерации текста может быть интерпретирован в русле теории порождающей грамматики как модель создания приемлемых текстов, обнаруживающих лингвосемиотическое сходство с антропологическими текстами. Это позволяет рассматривать сгенерированный текст как факт новой коммуникативной практики в рамках взаимодействия

человека и машины и изучать его как лингвистический феномен наряду с текстами, являющимися продуктом речевой деятельности человека.

1.4. Оптимизация и генерация как механизмы функционирования текстов в цифровой среде

Стратегия оптимизации текста как механизм улучшения поисковой выдачи источника при обращении к нему пользователя приобретает особое значение для текстов, функционирующих в цифровой среде. Термин «SEO», иначе «поисковая оптимизация» (Search Engine Optimization), появился в англоязычной литературе в середине 1990-х гг. для обозначения технологий работы с веб-ресурсами для журналистов. В русскоязычных лингвистических исследованиях представление «seo-текст» появилось позже [Кононова, Николаев 2016], хотя технологии оптимизации успешно применялись в практической деятельности журналистов и копирайтеров.

На сегодняшний день оптимизация выступает в качестве способа продвижения текста за счёт его приоритезации при поисковой выдаче в браузере в ответ на запрос пользователя. Природа оптимизации изначально обеспечивается технологиями, работающими на основе встроенных алгоритмов поисковых систем, которые способны ранжировать результаты поисковой выдачи по степени релевантности их содержания и структуры условиям, заданным в поисковом запросе.

На наш взгляд [Остапенко 2024], «структура seo-текста, помимо технических аспектов продвижения его содержания, связана с лингвистикой текста (его семантикой, синтактикой и прагматикой) в аспекте его адаптации под нужды адресата. Конечная цель такой адаптации состоит в повышении коммуникативной эффективности текста» [Остапенко 2024: 6].

Сервисы, предлагающие услуги по оптимизацию текстов, позволяют изменять структурно-содержательные характеристики исходного текста в следующих аспектах:

1. Трансформация семантического ядра текста за счет определения ключевых слов и их перераспределения на основе выявления плотности ключевых слов для каждого структурного компонента текста во избежания превышения допустимой доли ключевой лексики от общего количества слов.

2. Изменение структурно-композиционных особенностей. Это, в свою очередь, подразумевает «членение на абзацы, маркировку структуры заголовками и подзаголовками, использование списков; наличие гипертекстовых структур за счет «перелинковки» текста с помощью внутренних и внешних ссылок как инструментов навигации по содержанию; формальные параметры текста (количество слов, средняя длина предложений, процент так называемой «воды», то есть соотношение стоп-слов и семантически значимых лексических единиц, и проч.)» [Остапенко 2024: 5].

3. Приоритезация прагматической ценности текста за счёт выделения метаописательных единиц (тегов), которые позволяют более эффективно организовать отбор и ранжирование текстов определённой тематической или адресной направленности.

Поскольку указанные приемы имеют не только технический, но и лингвистический аспект функционирования, влияющий на семантику и прагматику текста процессы оптимизации, можно считать новой коммуникативной стратегией, которая определяет эффективность его функционирования в интернет-среде и влияет на востребованность текста среди других, сходных по тематике, ресурсов.

Прикладное использование инструментов оптимизации цифрового текста привело к появлению различных автоматизированных сервисов текстового анализа, позволяющих получать актуальную информацию о формально-семантических параметрах текста. Это, с одной стороны, привело к сокращению времени, необходимого для создания seo-текста, а с другой, обострила проблему выбора релевантных источников для пользователя, поскольку оптимизация текста не является синонимом его коммуникативного качества. Как показывает практика, seo-текст «удобен» для поисковых систем, при этом другие характеристики текста

(содержательность и доступность представленной информации, логичность изложения, эстетичность и проч.), значимые для пользователя, не входят в параметры оптимизации.

Мы считаем [Остапенко 2024], что автоматическая генерация текста является новой коммуникативной стратегией создания цифрового текста, функционирующего в современной интернет-среде. При этом, развивая аспекты формальной оптимизации и параметризации текста, на первый план выходит разработка компьютерных алгоритмов, способных генерировать связный текст, соответствующий заданным параметрам и характеристикам пользователя. Становясь частью коммуникативной практики, сгенерированный текст расширяет сферу функционирования, захватывая сферы новостного, рекламного, образовательного и развлекательного контента.

Хотя доля сгенерированного контента в общей массе электронных ресурсов, функционирующих в интернет-пространстве, растет, на сегодняшний день не существует объективных и достоверных данных, касающихся их приблизительного числа. Большинство исследований, связанных с этим вопросом, проводятся в рамках опросов общественного мнения¹, опирающихся на личный опыт использования респондентами нейросетевых алгоритмов для генерации текстового и визуального контента. Данные платформы «Антиплагиат» на начало 2024 года свидетельствуют о 20% таких текстов в общей доле академических работ². Целенаправленные исследования в этой области показывают³, что доля сгенерированного контента уже превышает 50%, при этом в поле

¹ Кто создает тексты и изображения с помощью ИИ? // Бюллетень «Цифровая экономика». Экспресс-информация. Национальный исследовательский университет «Высшая школа экономики». Институт статистической информации и экономики знаний [Электронный ресурс]. – Режим доступа: <https://issek.hse.ru/mirror/pubs/share/904851575.pdf> (Дата обращения: 28.07.2024).

Нейростат: Статистика знания и использования генеративных нейросетей // Поискковая система «Яндекс» [Электронный ресурс]. <https://ya.ru/ai/stat> (Дата обращения: 28.07.2024).

² «Антиплагиат»: каждая пятая работа российских студентов имеет следы ИИ // Интернет-портал «Сnews» [Электронный ресурс]. – Режим доступа: https://www.cnews.ru/news/line/2024-07-24_antiplagiat_kazhdaya_pyataya (Дата обращения: 28.07.2024).

³ Исследование: ИИ повлиял на более чем половину контента в интернете // Интернет-портал «4pda.to» [Электронный ресурс]. – Режим доступа: https://4pda.to/2024/09/15/432510/issledovanie_ii_povliyal_na_bole_chem_polovinu_kontenta_v_internete/ (Дата обращения: 15.09.2024).

исследовательского внимания попадают тексты на разных языках, в том числе полученные с помощью автоматического машинного перевода.

В связи с ростом доли сгенерированного текста в общей доле текстов на первый план выходят вопросы, связанные с формированием новых стратегий текстопорождения, становящиеся значимыми аспектами коммуникативной практики. Мы считаем, что к таким коммуникативным стратегиям можно отнести: «автоматизацию коммуникативно-речевых задач, связанных с созданием текстов типовой структуры (например, информационных или коммерческих); автоматическое размножение одного текста с целью получения множества текстовых произведений для различных каналов интернет-дискурса; параметризацию текста, который предстаёт как набор лингвистических и экстралингвистических характеристик, отвечающих целям текстопорождения; прогнозирование содержания и эффективности текста, который соответствует коммуникативным ожиданиям субъекта; формализацию процессов создания и редактирования текста, превращающую активного субъекта речевой деятельности в оператора программного алгоритма; относительную независимость содержания и качества текста от языкового и коммуникативного опыта субъекта, «надличный» характер, лишенный ярко выраженных индивидуально-авторских черт; возможности экспериментирования с содержанием и стилем текста, находящиеся вне зоны коммуникативной ответственности субъекта» [Остапенко 2024: 7].

Очевидно, что относительная независимость автоматической текстогенерации от человеческого фактора приводит к феномену отчуждения текста от речевой деятельности субъекта, что в целом соответствует тенденции к формированию постгуманистической парадигмы в познании. Традиционный текст предстаёт как процесс и продукт речевой деятельности, включающий три компонента: мотивационно-целевой, операциональный и результативный, а также обладает определённой прагматикой, обусловленной коммуникативными намерениями адресанта (автора). Сгенерированный текст не может быть осмыслен как продукт речевой деятельности, поскольку он изолирован от речевой деятельности субъекта и организован по определенным вероятностным моделям

генерации высказываний на естественном языке. При этом все компоненты речевой деятельности, присутствующие в антропологическом тексте, в сгенерированном тексте лишаются субъектно-деятельностной природы и становятся функцией диалогового агента. При этом вопрос о прагматическом статусе искусственно сгенерированного текста остается открытым: программный алгоритм не обладает коммуникативными интенциями, однако коммуникативный замысел, эксплицируемый в запросе (промте) на генерацию текста, по-прежнему принадлежит субъекту. Поэтому мы считаем, что «можно говорить о появлении текстовых феноменов с редуцированной речедеятельностной составляющей, имитирующих свойства речевых высказываний, созданных человеком» считает [Остапенко 2024: 8]. Отчуждение текста от субъекта речемыслевого процесса может быть рассмотрено с различных точек зрения, в том числе и в аксиологическом аспекте как феномен нового этапа эры информационного потребления. Текст, становясь источником фактологической информации (данные⁴) и средством монетизации интернет-дискурса, лишается информационно-культурной ценности и творческой индивидуальности.

Таким образом, проблема текста, сгенерированного нейросетевыми алгоритмами, является не только одной из проблем теории информации, взятой в аспекте соотношения данных и информации, но и выступает как более общая культурно-лингвистическая проблема. Вопрос о содержательной ценности сгенерированного текста остается открытым. С одной стороны, его информативность отрицается, поскольку в процессе генерации не происходит приращения знаний: алгоритмы компилируют текстовые данные, полученные на основе больших обучающих выборок, по заданным параметрам, ориентируясь

⁴ Вопросам определения понятия «информация» посвящены классические исследования по кибернетике с 1940-х годов. Разработкой теории информации в количественном и качественном аспектах в разное время занимались Н. Винер, У. Р. Эшби, К. Э. Шеннон, Р. Хартли, Дж. фон Нейман, У. Бар-Хиллел и Р. Карнап. В рамках теории информации были разграничены понятия «данные» и «информация». Данные представляют собой набор сведений об объектах, явлениях, фактах окружающей реальности, в то время как информация представляет собой данные, подвергнутые структурированию, организации и когнитивной обработке путем включения в контекст человеческого знания и деятельности.

прежде всего на грамматическую и смысловую связность текста, его соответствие нормам естественной языка. При этом процесс генерации текста изолирован от когнитивных процессов, которые сопутствуют речепорождению субъекта. Отсутствие когнитивного «опыта» порождает фактические ошибки, ложные и ошибочные сведения в процессе генерации и требует от оператора нейросетевых алгоритмов постоянного фактчекинга, направленного на установление связи между содержанием текста, созданного алгоритмом, и внеязыковой действительностью. Мы полагаем, что в терминах парадигмы постмодернизма феномен сгенерированного текста превращается в текстоид, квазитекст, или симулякр, знак без означаемого, формируемого индивидуальным опытом активного субъекта <...> Сгенерированный текст воспроизводит только означающее, эквивалентное грамматической и смысловой связности на уровне языковой и композиционной структуры, выступая механизмом оптимальной компиляции коммуникативно приемлемых отрезков сообщений в текстовое единство» [Остапенко 2024: 9].

Искусственный текст является вызовом не только для лингвистики, но и для коммуникативной практики в целом. Ограниченность его функционирования взаимодействием человека и диалоговых агентов (чат-боты, голосовые ассистенты, системы «умного» поиска) постепенно сменяется экспансией сгенерированного текста в сферу маркетинга, медиакоммуникаций, образования, создания цифровых двойников личности. Эти процессы требуют не только правовой и этической оценки, но и исследований с точки зрения лингвистики текста, позволяющих определить коммуникативную природу данного феномена в аспекте сходства и различий с антропологическими текстами.

1.5. Методология анализа сгенерированного текста и оценка его качества

Изучение сгенерированных текстов как особого типа неструктурированных больших данных (big data) актуализирует значимость автоматических методов обработки естественных языка с помощью инструментов компьютерной

лингвистики. Будучи разновидностью формального языкового анализа, методы компьютерной лингвистики позволяют разрабатывать алгоритмы для автоматизации основных процессов, связанных с автоматической генерацией и интерпретацией сообщений естественного языка, включая решение таких задач, как тематическая классификацией сообщений, анализ тональности высказывания, извлечение информации, устранение контекстуальной неоднозначности, создание цифровых корпусов текстов, распознавание звучащей речи и других.

Наиболее перспективными инструментами формального автоматического анализа применительно к сгенерированным текстам выступают методики статической оценки, модели машинного обучения, в том числе нейронные сети, способные обучаться на больших объемах текстовых данных, а также методы векторной математики, создающие модели плотных векторных представлений слов в корпусе текстов на основе контекста для установления семантической близости или решения вопроса о классификации лексико-семантических единиц [Автоматическая обработка текстов... 2017].

Для лингвистического анализа сгенерированного текста методы компьютерного анализа выступают инструментом количественной оценки параметров текста, позволяют обнаруживать статистически значимые закономерности в дистрибуции языковых единиц, а также аппроксимировать неочевидные зависимости между формальными характеристиками текстов. Следует отметить, что анализ текста, созданного с помощью нейросетевых алгоритмов, лингвистическими методами не предполагает решения инженерных задач, связанных с построением архитектуры нейронных сетей, поскольку предметом такого анализа выступает текст на естественном языке. В связи с этим методы автоматической обработки выступают как вспомогательные процедуры количественной и качественной оценки данных, полученных в результате анализа формально-содержательных особенностей такого текста.

Наиболее ценными для лингвистических исследований являются автоматические методы анализа семантики сгенерированных текстов, позволяющие определить смысловую структуру текста, устанавливая

семантические роли, связи и отношения, существующие между ключевыми компонентами его содержания, выявлять именованные сущности и осуществлять смысловое сжатие (реферирование) текста. К автоматическим методам анализа семантики примыкает и тематическое моделирование, позволяющее на основе вероятностного распределения слов в тексте выявлять латентные (скрытые) темы.

Рассмотрим, как методы автоматической обработки текста могут быть использованы при работе со сгенерированным текстом.

Сгенерированный текст является структурным образованием, включающим несколько уровней анализа: фонологический (для звучащей речи) и графо-символический (для письменных сообщений), морфолого-синтаксический, семантический, дискурсивно-прагматический, а также уровень обработки текста с помощью программных инструментов [Бочаров, Митренина, Добров 2017].

Обработка текста предполагает извлечение информации и её трансформацию в машиночитаемый формат для последующего решения задач, связанных с классификацией, анализом и генерацией текстовых данных. В качестве базовой единицы извлечения текстовой информации выступает символ (буква, цифра и знак препинания). Кодировка символов, например, UTF-8, определяет их представление в цифровом формате.

Минимальными смысловыми единицами на этапе обработки данных текста являются токены, обычно равнозначные отдельным символам – графемам, знакам препинания или другим значимым единицам текста. Токенизация – это процесс разделения текста на токены, позволяющий упрощать дальнейший анализ и переходить к более сложным единицам высших уровней – предложениям, текстам и их корпусам.

Предложение в компьютерной лингвистике равнозначно группе токенов, которые образуют законченное высказывание. Визуальными границами предложений, которые выделяются с помощью комбинированных методов автоматической обработки текста, выступают фрагменты текста, начинающиеся с заглавной буквы и заканчивающиеся знаками завершения. Синтаксический анализ применительно к полученным наборам токенов – высказываний предлагает

способы структурирования предложений, включая их разбивку на грамматические элементы.

Параграфы и тексты – более крупные единицы, содержащие множество предложений и представляющие собой связные смысловые фрагменты.

Наконец, корпус текстов – собранный и организованный набор текстов, который используется для анализа, разработки и оценки языковых моделей и инструментов. Корпус может включать в себя тексты различных жанров, стилей, тем и источников, и его основной целью является предоставление репрезентативного образца языка (в его узуальном употреблении) для научных исследований и практического применения.

Основной метод генерации текста – компьютерное моделирование, которое включает процедуры предварительной обработки обучающих текстовых данных, собственно моделирования, обучения и оценки результатов [Бенгфорт, Билбро, Охеда 2019].

Первый шаг в компьютерном моделировании текста – токенизация (рис. 1), которая позволяет создавать n-граммы или векторные представления (эмбединги).



```
# 1. Токенизация
def tokenize(text):
    tokens = nltk.word_tokenize(text, language='russian')
    return tokens

tokens = tokenize("Токенизация текста - процесс разбивки его на отдельные элементы: символы, слова и знаки препинания.")
for token in tokens:
    print(token)
```

Токенизация
текста
-
процесс
разбивки
его
на
отдельные
элементы
:
символы
,
слова
и
знаки
препинания
.

Рисунок 1. Токенизация текста с использованием NLTK

Инструменты токенизации являются достаточно гибкими, позволяя реализовывать как автоматические, так и пользовательские стратегии разбиения текста на элементы для последующего анализа. Сложность токенизации связана с неравнозначностью слова и его графической репрезентации. Например, автоматическая токенизация последовательностей вида «И. И. Иванов», «куда-либо», «Чёрное море» затруднена в связи с необходимостью выделения в качестве токенов не отдельных символов ('И', '.', 'Иванов', 'куда', '-', 'либо', 'Чёрное', 'море'), а их сочетаний, эквивалентных по функционалу одному слову. Данная задача может быть решена либо использованием пользовательских регулярных выражений [Гречачин 2016], либо созданием словарей исключений, включающих составные наименования, встречающиеся в данном тексте, либо более продвинутыми методами машинного обучения, в том числе с использованием современных нейросетевых моделей [Бартеньев 2023].

Токенизация текста связан с процессом создания N-грамм — последовательностей из n токенов, которые используются для моделирования текста [Гудков, Гудкова 2011]. Например, биграммы (2-граммы) и триграммы (3-граммы) представляют собой последовательности из двух и трех токенов соответственно. N-граммы используются для предсказания вероятности появления слова в контексте, а также для определения ключевых паттернов (тематических групп) текста (рис. 2).

```
# Моделирование текста с помощью n-грамм
tfidf_vectors, tfidf_features = tfidf_vectorizer(cleaned_text, ngram_range=(1, 3))
print("TF-IDF n-граммы:", tfidf_features)
```

TF-IDF n-граммы: ['векторных' 'векторных представлений' 'векторных представлений слов' 'единицами' 'единицами текста' 'единицами текста необходимыми' 'знаки' 'знаки препинания' 'знаки препинания токены' 'минимальными' 'минимальными смысловыми' 'минимальными смысловыми единицами' 'необходимыми' 'необходимыми построения' 'необходимыми построения эмбедингов' 'отдельные' 'отдельные элементы' 'отдельные элементы символы' 'построения' 'построения эмбедингов' 'построения эмбедингов векторных' 'представлений' 'представлений слов' 'препинания' 'препинания токены' 'препинания токены являются' 'процесс' 'процесс разбивки' 'процесс разбивки отдельные' 'разбивки' 'разбивки отдельные' 'разбивки отдельные элементы' 'символы' 'символы слова' 'символы слова знаки' 'слов' 'слова' 'слова знаки' 'слова знаки препинания' 'смысловыми' 'смысловыми единицами' 'смысловыми единицами текста' 'текста' 'текста необходимыми' 'текста необходимыми построения' 'текста процесс' 'текста процесс разбивки' 'токенизация' 'токенизация текста' 'токенизация текста процесс' 'токены' 'токены являются' 'токены являются минимальными' 'элементы' 'элементы символы' 'элементы символы слова' 'эмбедингов' 'эмбедингов векторных' 'эмбедингов векторных представлений' 'являются' 'являются минимальными' 'являются минимальными смысловыми']

Рисунок 2. Моделирование текста с использованием TF-IDF на основе n-грамм

Для компьютерной лингвистики актуальной является процедура нормализации текста, позволяющая привести текст к стандартному виду для последующих преобразований. Унификация текста производится на формально-содержательном уровне. Сначала все слова текста приводятся к одному регистру, из текста удаляются знаки препинания, небуквенные символы. Далее производится очистка, позволяющая исключить из текста неполнозначные слова (предлоги, союзы, частицы) и другие элементы, отнесённые к категории стоп-слов. Содержательная унификация связана со стандартизацией грамматической формы слов. Для этого используют процедуру лемматизации – приведение слов к их базовой или словарной форме (лемме). Например, слова «бегаю», «бегал» и «бег» приводятся к одной лемме «бежать». Лемматизация отличается от токенизации тем, что она направлена на нормализацию формы слов, тогда как токенизация просто разделяет текст на составляющие. Лемматизация обычно происходит после токенизации, чтобы сгруппировать различные формы одного и того же слова [Bird, Klein, Loper 2009; Большакова, Клышинский, Ландэ 2011].

Наряду с лемматизацией применяется и стемматизация – отсечение флексий (окончаний и части формообразующих аффиксов) для получения основы слова. Полученные в результате преобразования стеммы имеют формальный характер, так как отсечение флексий производится без учёта морфологических особенностей слова. В связи с этим лемматизация, несмотря на временную затратность, является более приемлемым методом обработки текста, так как улучшает точность анализа, особенно в задачах, требующих понимания значений слов (например, для проведения семантического анализа высказывания или для извлечения информации) (рис. 3).

Лемма	Стемма
токенизация	токенизац
текст	текст
-	-
процесс	процесс
разбивка	разбивк
его	ег
на	на
отдельный	отдельн
элемент	элемент
:	:
символ	символ
,	,
слово	слов
и	и
знак	знак
препинание	препинан
.	.

Рисунок 3. Результаты лемматизации и стемматизации текста

Для обеспечения работы компьютера с текстовыми данными производится представление слов в виде векторов чисел, иначе называемых эмбедингами. Векторная кодировка текста является отличительной особенностью его машинной обработки, не применяемой при анализе текста человеком. Векторные представления слов в многомерном пространстве производятся автоматически с помощью готовых инструментов векторизации, которые захватывают семантические и синтаксические особенности слов (рис. 4).

	TF-IDF
текста	0.408248
векторных	0.204124
символы	0.204124
эмбедингов	0.204124
элементы	0.204124
токены	0.204124
токенизация	0.204124
смысловыми	0.204124
слова	0.204124
слов	0.204124
разбивки	0.204124
единицами	0.204124
процесс	0.204124
препинания	0.204124
представлений	0.204124
построения	0.204124
отдельные	0.204124
необходимыми	0.204124
минимальными	0.204124
знаки	0.204124
являются	0.204124

Рисунок 4. Результат создания векторных представлений слов

Используя формальные показатели (например, косинусное расстояние между векторами), можно определить степень семантической близости слов,

охарактеризовать положение слова относительно других слов в предложении и контекст его употребления [Батраева, Нарцев, Лезгян 2020].

Помимо методов машинного анализа, изучение сгенерированных текстов монологической и диалогической природы может опираться на традиционные подходы, используемые в языкознании. Так, объектом лингвистической рецепции может быть структурно-семантическая природа сгенерированного текста, его коммуникативно-речевая или языковая составляющие. В зависимости от целей анализа в исследовании сгенерированных текстов большого объема и (или) количества применяются методы тематического моделирования с использованием алгоритмов машинного обучения, автоматического анализа морфолого-синтаксической структуры текста, определения эмоциональной окраски (тональности) текстов.

Продуктивным является сочетание количественных и качественных методов исследования, которое позволяет устанавливать статистические закономерности в текстовых данных и интерпретировать их с опорой на теоретико-методологическую базу лингвистических исследований, посвященных анализу текста и дискурса.

Поскольку сгенерированный текст является новым для лингвистической науки феноменом рецепции, в настоящее время активно исследуется возможность применения к его анализу традиционных лингвистических методов и приёмов, а также определяется критерии ограничения тех или иных методических подходов. Так, некоторые исследователи считают прагматический анализ сгенерированного текста [Черкасова, Тактарова 2024] нецелесообразным и даже невозможным, поскольку алгоритмы на основе больших языковых моделей не обладают собственными коммуникативными установками, не будучи субъектами коммуникативной ситуации. Однако такой подход к сгенерированному тексту не учитывает его неотделимость от речевого задания – промта (запроса), который исходит от антропологического субъекта для решения определённых коммуникативных задач. Прагматика сгенерированного текста, который создан языковой моделью, опосредована промтом, поэтому, с нашей точки зрения,

методологически корректным является рассмотрение сгенерированного текста в динамическом аспекте – в процессе текстопорождения и в контексте формирующего его запроса. Динамическая природа сгенерированного текста, его структурная и содержательная связь с запросом позволяет исследовать сгенерированные сообщения как коммуникативные единицы новой природы, возникающие в процессе взаимодействия антропологического субъекта и диалогического агента на основе искусственного интеллекта.

Оценка коммуникативных качеств автоматически сгенерированного текста выступает как междисциплинарная проблема, поскольку генераторы на основе нейронных сетей представляют собой программные алгоритмы со встроенными системами оптимизации функций потерь при вычислении. Соответственно оценка точности работы модели определяется критерием минимизации ошибок вычислений, который для лингвистики не имеет характер информативного интерпретируемого критерия коммуникативного качества сгенерированного текста. Поясним это на примере. Текст, созданный на естественном языке, согласно рассмотренной нами концепции Н. Хомского, должен отличаться лингвистической приемлемостью как совокупностью трех факторов (семантическая, грамматическая и прагматическая приемлемость) для того, чтобы быть воспринятым как адекватный и системе языка, и коммуникативным ожиданиям адресата. Аномальный с грамматической, семантической или прагматической точки зрения текст не может считаться лингвистически приемлемым для типовых ситуаций коммуникации, при этом модель может посчитать коммуникативные качества такого текста относительно удовлетворительными, поскольку ориентируется не на контекст коммуникации, а на совокупность параметров, заданных её архитектурой. Поэтому встает вопрос о поиске и подборе собственно лингвистических критериев оценки коммуникативных качеств сгенерированного текста. Это поле является проблемным для исследования, так как нуждается в определении набора минимальных критериев, обеспечивающих лингвистическую приемлемость сгенерированного текста, а также поиска универсальных критериев для оценки

коммуникативно-речевых качеств сгенерированного текста разных жанров и стилей.

Понятие «коммуникативное качество текста» применительно к процессам генерации может быть интерпретировано в культурно-речевом, эстетическом или этическом аспектах.

Традиционные коммуникативно-речевые критерии качества опираются на принципы точности, полноты, логичности, правильности высказывания, его соответствия коммуникативной ситуации, богатства словаря, разнообразия речевых конструкций, выразительности, чистоты, уместности речевого произведения [Черкасова, Тактарова 2024: 2227–2231]. Данные критерии могут быть применены к оценке коммуникативных качеств сгенерированных текстов, поскольку такие тексты создаются и функционируют наряду с антропологическими текстами в коммуникативной практике.

В эстетическом аспекте вопрос качества сгенерированного текста перемещается в плоскость оценки его художественных достоинств, красоты стиля, уникальности используемых приёмов выражения мысли. Очевидно, что эстетические требования качества выходят за рамки собственно лингвистических и приобретают статус культурно-исторических критериев, обеспечивающих ценность текста как факта искусства слова. При этом эстетический критерий становится значимым для результатов текстогенерации, которым придаётся статус «художественных произведений», созданных с помощью инструментов автоматической генерации текста [Кучина 2017].

Наконец, этический критерий качества как соблюдение морально-нравственных норм, закреплённых в коммуникативной традиции, накладывает ограничение на генерацию текстов, нарушающих авторские права, фальсифицирующих академические результаты, формирующих у субъекта искажённый облик реальности, содержащих недопустимые оценки (например, этнический, религиозные, политические), артефакты языковой дискриминации.

Трудности в выборе лингвистических критериев коммуникативной эффективности сгенерированного текста также связаны с необходимостью

построения универсальных инструментов оценки, которые должны быть адекватны текстам разных стилей и жанров. В то же время жанрово-стилевые различия текстов зачастую обуславливают характер оценки коммуникативных качеств текста. Так, например, канцеляризм уместен в текстах официально-делового стиля, но их присутствие в текстах других стилей служит маркером речевой ошибки и, как следствие, снижает коммуникативно-речевое качество текста. Поэтому поиск универсальных лингвистических критериев оценки сгенерированного текста естественным образом ограничен самой природой исследуемого языкового объекта.

Несмотря на трудности, связанные с построением лингвистической модели оценки качества сгенерированного текста, исследователи занимаются поиском таких критериев, рассматривая разные типы конечных продуктов процесса текстогенерации. Наиболее подробно данная проблема разработана в отношении автореферирования текста, поскольку в этот процесс включен исходный текст, выступающий в виде эталона оценки. Так, ряд исследователей оценки процесса и результата автореферирования текста обращают внимание на такие критерии качества полученного реферата, как содержательное и структурное соответствие исходному тексту [Белогорская, Резанова 2024]. При этом наиболее частым недочетом, снижающим качество автореферата, сгенерированного программными средствами, выступает отсутствие четких логических связей (переходов) между частями текста.

Проблема коммуникативного качества сгенерированного текста не может быть рассмотрена в отрыве от его соотношения с содержанием запроса (промта), поскольку именно соответствие сгенерированного текста параметрам, заданным в запросе, определяет дальнейшее коммуникативное поведение субъекта: если оператор считает сгенерированный текст удовлетворительным, он использует его для тех целей, для которых он был создан; в противном случае субъект либо генерирует текст повторно, видоизменяя запрос, либо подвергает редактированию уже имеющийся текст, желая привести его в соответствие своим коммуникативным целям и ожиданиям.

Таким образом, коммуникативная оценка сгенерированного текста может быть рассмотрена как процесс, который осуществляется в двух аспектах:

1. Объективная оценка, определяющая степень грамматической и семантической приемлемости текста, его соответствие коммуникативно-речевым качествам и нормам.

2. Субъективная оценка, включающая релевантность созданного текста коммуникативным целям и ожиданиям субъекта (его функциям, речевой прагматике).

Очевидно, что качественный с коммуникативной точки зрения сгенерированный текст соответствует двум параметрам оценки, выступая и как приемлемый с точки зрения грамматики и семантики, и как коммуникативно эффективный (прагматически приемлемый).

Оценка качества сгенерированного текста тесно связана с проблемой разграничения искусственного и естественного текста. Иными словами, данный аспект исследований связан с обучением нейросетевых алгоритмов механизмам определения лингвопрагматической приемлемости сгенерированного ими текста на основании его соотнесения с коммуникативными паттернами текстов, созданных человеком. В настоящее время данная проблема разрабатывается в русле проекта RuCoLA⁵ (Russian Corpus of Linguistic Acceptability), который представляет собой большой набор структурированных текстовых данных, размеченных по бинарной шкале приемлемости. Данные образуют русский корпус приемлемости, включающий 13,4 тысячи предложений с бинарными оценками приемлемости и содержащий подробное описание четырех категорий нарушений для 3,7 тыс. неприемлемых предложений. RuCoLA охватывает два типа источников данных: лингвистическая литература (корпусная лингвистика) и предложения, созданные с помощью генеративных моделей. В связи с этим возникает возможность сопоставлять искусственные и естественные тексты в аспекте их сходств и различий в терминах лингвистической приемлемости.

⁵ Инициаторами проекта выступили исследователи и NLP-разработчики из SberDevices, ABBYY, Yandex Research, Huawei Noah's Ark Lab и Факультета компьютерных наук ВШЭ.

Понятие «приемлемости», используемое как метка для бинарной классификации высказываний корпуса, восходит к идеям Н. Хомского, который оценивал приемлемое предложение как «совершенно естественное», «сразу понятное» и «никоим образом не причудливое или диковинное» с точки зрения носителя языка (субъекта) [Chomsky 1957]. Понятие приемлемости, по мнению лингвиста, градуальное, то есть имеет свои степени, зависящие от различных факторов (например, быстроты, правильности и единообразия запоминания и распознавания или нормальности интонации). Также приемлемость выступает явлением комплексным, сочетающим лингвистическую семантику, синтактику и прагматику, поскольку грамматически приемлемое высказывание не может считаться лингвистически приемлемым, если оно семантически аномально или ограничено в плане запоминания из-за большого объема, интонационных или стилистических факторов.

В результате проведенных исследований, а также обучения моделей распознаванию сгенерированных текстов, не отвечающих критериям приемлемости, были выявлены лингвистические факторы, относящиеся к разным уровням языковой системы и влияющие на качество генерируемого текста:

на уровне синтаксиса: нарушения согласования, нарушение порядка слов, неправильное построение синтаксических конструкций, неправильное использование приложений (аппозиций), нарушения в использовании переходных глаголов, а также в количестве и типе аргументов глагола (агенс, пациенс, адресат, инструмент, локатив), пропущенные грамматические элементы;

на уровне морфологии: неправильное словообразование или словоизменение, ошибки в образовании грамматических форм, появление несуществующих слов;

на уровне семантики: неправильное использование отрицания, нарушение семантической структуры аргументов глагола;

лингвистические аномалии на уровне формы и содержания (так называемые «галлюцинации»): искажение текста, бессмысленные предложения, неуместные повторы.

Приведенные примеры показывают, что лингвистическая приемлемость в целом не решает проблему оценки коммуникативного качества сгенерированного текста. Так, например, наличие грамматических аномалий хотя и свидетельствует о нарушении грамматических норм языка, зачастую происходит в рамках возможностей языковой системы, поэтому грамматические аномалии нередки в естественных текстах и, как правило, не вызывают специфических затруднений в понимании высказывания. В то же время избыточная «правильность» речи может служить признаком генерации текста. Тем не менее указанные критерии качественного (лингвистически приемлемого) текста являются более надежными для лингвиста, чем количественные показатели, поскольку они легко интерпретируются, соотносятся с традиционными аспектами лингвистического анализа текста и позволяют получать масштабируемые данные применительно к имеющимся у исследователя коллекциям текстовых данных.

Несмотря на сложность и многоаспектность проблемы оценки коммуникативных качеств сгенерированного текста, можно отметить, что на сегодняшний момент инициированы исследования, которые позволяют выделить отдельные критерии такой оценки, отличающиеся надежностью и объективностью. Тем не менее проблема текстогенерации, не являясь сугубо лингвистической, задействует и ряд экстралингвистических факторов, касающихся соответствия содержания сгенерированного текста реальности (оценки его фактологической достоверности), соблюдения этических норм коммуникации, коммуникативного эффекта, производимого текстом на субъекта (субъективная удовлетворенность или неудовлетворенность результатами работы с нейросетью-генератором текстового контента). В связи с этим решение проблемы оценки качества сгенерированного нейросетью текста находится в области градуальной оценочной парадигмы (от менее приемлемого к более) и требует разработки дальнейшего совершенствования лингвистической методологии анализа такого текста, учитывающей его специфику в аспекте цифровизации и автоматизации лингвистических исследований.

Выводы по первой главе

Современная лингвистика активно исследует сгенерированные тексты, в иной терминологии – искусственные тексты, которые возникают в условиях цифровой коммуникации и характеризуются интерактивностью, гипертекстуальностью, мультимедийностью и адаптивностью.

Анализ лингвистических основ генерационных процессов позволяет обнаружить, что методологической базой для исследования сгенерированного текста являются лингвистические концепции, базирующиеся на идеях трансформационной грамматики, динамического порождения текста и дериватологической теории текстообразования как механизма варьирования текстовых структур на основе заданного инварианта.

Автоматическая генерация текстов и их специальная поисковая оптимизация обеспечивают новые возможности для создания контента в сфере новостей, рекламы, образования. Экспансия сгенерированного текста поднимает вопрос содержательной ценности автоматически создаваемых сообщений и их способности передавать информацию, сохраняя при этом структурные нормы языка. Эта проблема многогранна и требует комплексного подхода, учитывающего результаты работы нейросетей в области грамматической и смысловой связности генерируемого сообщения и отражения в его содержании когнитивного опыта, которым не обладают компьютерные алгоритмы.

Появление сгенерированных текстов, созданных с помощью нейросетей, требует новой методологии анализа, поскольку они представляют собой комбинацию языкового материала и алгоритмических процессов на основе заданных параметров с минимальным участием антропологического субъекта. Параметры анализа такого текста включают морфолого-синтаксический и семантический уровни исследования, а также использование машинного обучения. Методы компьютерной обработки естественного языка автоматизируют процессы анализа сгенерированных текстов, обеспечивая их перевод в машиночитаемый формат. Процесс включает приведение слов к единому

регистру и удаление знаков препинания и стоп-слов. Лемматизация и стемматизация используется для нормализации формы слов, позволяя сгруппировать различные формы одного и того же слова. Векторы чисел, или эмбединги, позволяют представить слова в виде векторных кодов, с которыми может работать компьютерный алгоритм. Наряду с автоматическими методами формально-статистического анализа сгенерированного текста применяются традиционные лингвистические методы, включающие тематический, лексико-семантический, структурно-композиционный, коммуникативно-речевой и иные виды лингвистического анализа, позволяющие провести качественное исследование сгенерированного текста.

Изучение сгенерированных текстов требует установления критериев оценки их коммуникативных качеств, что, в свою очередь, является междисциплинарной задачей, поскольку генерация на основе нейронных сетей имеет алгоритмические особенности. Определение качества языкового моделирования основывается на критериях минимизации ошибок, которые могут оказаться неинформативными для лингвистики. Чтобы текст считался коммуникативно эффективным, он должен соответствовать критериям семантической, грамматической и прагматической приемлемости. Проблемным полем в этой сфере является разработка собственно лингвистических критериев для оценки качества сгенерированного текста, применимых для продуктов генерации различных жанров и стилей. Качество текста можно рассматривать не только в лингвистическом, но и в культурно-эстетическом аспекте с точки зрения господствующего речевого идеала и коммуникативных ожиданий пользователей. Многообразие факторов, определяющих критерии оценки коммуникативного качества сгенерированных текстов, а также относительность данных критериев, формирует градуальный подход к оцениванию искусственно созданных текстов и требует разработки методологических подходов к их атрибуции и анализу.

ГЛАВА 2. СТРАТЕГИИ ИНФОРМАЦИОННОГО ТЕКСТОПОРОЖДЕНИЯ: КОММУНИКАЦИЯ С ДИАЛОГОВЫМИ АГЕНТАМИ

2.1. Компьютерно-опосредованная коммуникация и диалоговая лингвистика

Компьютерно-опосредованная коммуникация, сформировавшаяся в эпоху дигитализации социальных интеракций и опосредованная современными компьютерными технологиями, стала предметом лингвистической рецепции в конце XX – начале XXI вв. [Баркович 2008], [Щипицина 2010]. Стремительное развитие компьютерных технологий обусловило разрыв между скоростью теоретического осмысления новых процессов взаимодействия в компьютерно-опосредованной среде и реальными трансформациями коммуникативной ситуации, связанной со стремительным развитием систем, основанных на генеративных алгоритмах. На смену антропологической коммуникации, опосредованной техническими аспектами, приходит гибридная коммуникация, субъектом которой становится компьютерный алгоритм, приобретающий персонифицированные черты.

В условиях экспансии «интернета вещей» (IoT) границы между человеческим и машинным взаимодействием становятся всё более размытыми, поэтому актуальным направлением исследования становится коммуникативная ситуация взаимодействия человека и диалогового агента. К наиболее перспективным направлениям исследования этой формы коммуникации человека и программного алгоритма можно отнести изучение лингвопрагматических аспектов речевого взаимодействия в рамках поиска форм успешной коммуникации (С. В. Киселева, А. А. Смирнова, Н. А. Трофимова, К. Е. Пластун, Н.Н. Зильберман, А. В. Копытова), а также практическое применение диалогических агентов в лингводидактических целях (А. П. Авраменко, С. В. Зенкина и другие).

Значимым аспектом исследуемого феномена является его анализ в русле диалоговой лингвистики, которую интересуют вопросы использования языка в условиях коммуникации между человеком и диалоговым агентом на базе технологий искусственного интеллекта [Шпильная 2018]. По мнению [Шпильная 2018], основой человеко-компьютерной коммуникации выступает принцип суппозиционной связи лексемы и текста, который предполагает развертывание потенциально заложенных в элементах текстовых сообщений смыслов через их условно-гипотетическую связь с контекстом коммуникативной ситуации. Применительно к антропологическому субъекту коммуникации это означает, что механизмы восприятия и интерпретации сообщений, генерируемых диалоговыми агентами, полностью определяется коммуникативным опытом и ожиданиями пользователя, который становится активным участником процессов смыслообразования.

Мы считаем важным отметить, что суппозиционный подход к восприятию и пониманию генерируемых в человеко-машинной коммуникации сообщений, должен сочетаться с признанием двуединой природы диалогового агента – технической и коммуникативной. С технической точки зрения, диалоговый агент, который может выступать в разных коммуникационных форматах (чат-бот, голосовой ассистент, интеллектуальная поисковая система), представляет собой коммуникативный интерфейс программного алгоритма – системы, с помощью которой организуется взаимодействие субъекта и машинного агента (например, нейронной сети). Однако диалоговые агенты представляют собой не только технические системы, но и коммуникативные агенты, которые функционируют в рамках «гуманизированных» стратегий текстопорождения.

Коммуникативный аспект функционирования диалогового агента изначально был связан с его функциональной ролью как средства автоматизации маркетинговых взаимодействий, участниками которых выступают компании (поставщики товаров и услуг) и целевая аудитория (клиенты – потребители товаров и услуг) [Козорис 2019]. Значительная коммерческая составляющая в использовании таких диалоговых систем акцентирует внимание на исследовании

критериев успешности коммуникативных ситуаций «человек – компьютерный алгоритм», в то время как исследованиям в области диалоговой природы дискурса с участием ИИ-агентов, а также механизмам порождения смысла с использованием сгенерированных текстов уделяется меньше внимания.

Для анализа коммуникации в системе «человек — диалоговый агент» в аспекте феномена текстогенерации актуальны следующие направления исследования:

1. Выявление особенностей построения речевых актов в коммуникации с диалоговым агентом, определение их типологических черт, независимых от специфики используемого языка.

2. Механизмы «приближения» уровня «речевой компетенции» диалогового агента к человеческому.

3. Корреляции речевых средств с признаками и параметрами дискурса, исследование маркеров социокультурной, прагматической, психологической, идеологической реальности в коммуникации.

4. Изучение процессов порождения и структурирования речевого диалогического высказывания, в основе которых лежат факторы, обуславливающие как ход речемыслительных и речеповеденческих процессов, так и конкретные формы возникающего при этом текстового диалогического продукта.

Перспективность исследования диалоговой природы коммуникации с использованием персонифицированного диалогового агента (чат-бота) [Зильберман 2009] подтверждается не только данными лингвистического анализа, но и изначальной установкой при моделировании работы таких систем на генерацию сообщений в ответ на запрос собеседника. Возможные варианты коммуникативно-речевых действий диалогового агента в ситуации диалога не являются свободными: они предопределены скриптами, которые представляют собой алгоритмы, содержащие определенные последовательности действий в ответ на запрос пользователя. Диалоговые скрипты предстают как технический аспект взаимодействия модели и пользователя. В терминах лингвистики процесс

моделирования речевых действий диалогового агента описывается через сценарий диалогов системы и пользователя.

Сценарий диалога для компьютерного агента представляет собой описание возможных вариантов развития коммуникативной ситуации, включая вероятные пользовательские запросы, ответы системы и переходы между различными состояниями диалога. Сценарий диалога чаще всего представляют в виде таблицы или «дерева решений», где узловым точкам соответствуют варианты выбора коммуникативных действий или тематики диалога.

Диалог ИИ-агента с пользователем может иметь следующую структуру:

1. Приветствие: диалоговый агент приветствует пользователя и предлагает выбрать услугу.
2. Варианты пользовательских запросов: получение общей информации о компании, получение об услугах или товарах, назначение даты и времени визита клиента.
3. Ответ диалогового агента на запрос клиента.
4. Подтверждение выбора пользователя.
5. Завершение коммуникации.

Данная структура может быть представлена более подробно в виде «дерева решений» (рис. 5).

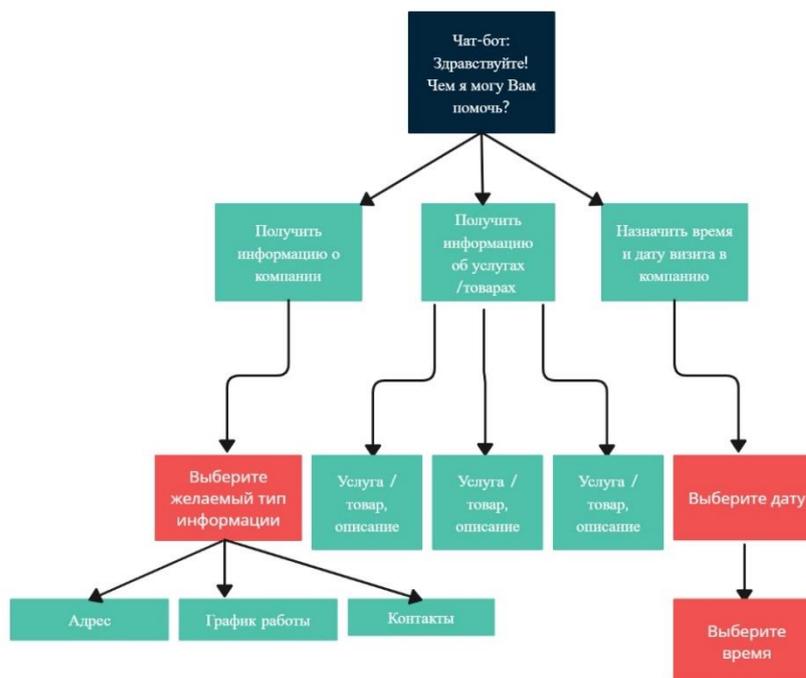


Рисунок 5. Схема диалога коммуникации с диалоговым агентом сценарного типа

В зависимости от выбора темы запроса, который осуществляет пользователь, диалоговый агент предлагает разные варианты ответа, соответствующие определённой ветке диалога. Например, ветка диалога для тематического запроса «Назначить время и дату визита в компанию» может включать сценарные реплики, касающиеся предложения выбрать тему для беседы из предложенного ботом списка, указать желаемую дату и время посещения компании, а также заключительные реплики, подтверждающие запись на прием и уточняющие наличие у клиента (пользователя) вопросов, не решённых в процессе чат-бот коммуникации.

Представленная сценарная ветка обладает логико-семантической схемой, включающей компоненты «субъект», в качестве которого выступает пользователь (U), предикат «приветствие» (G) и объект – помощь (консультация), предложенная ботом (H).

Иные компоненты схемы представлены в логико-символической нотации:

$I(S, N) \wedge N(\text{сведения}) \rightarrow P(\text{информация о компании})$ – «Компания находится [адрес], [график работы], [контакты]», где субъект – компания (S), предикат – информация (I), объект – запрашиваемые сведения (N).

$R(U, A) \wedge T(\text{дата}) \rightarrow P(\text{записаться})$ – «Вы записались на прием [дата] [время]», где субъект – пользователь (U), предикат – записаться (R), объект – прием (A).

$O(S, U) \wedge U(\text{услуги}) \rightarrow P(\text{услуга})$ – «Вы выбрали услугу [услуга]», где субъект – компания (S), предикат – услуги (O), объект – услуги (U).

Нетрудно заметить, что отдельные ветки диалога идентичны в формально-логическом аспекте. Ответ диалогового агента всякий раз формируется как формально-логическая операция импликации двух пропозиций, которые заданы в сценарии диалога изначально. В свою очередь, связь между логическим выражением и пропозициональной структурой диалога пользователя и диалогового агента заключается в том, что логическое выражение используется для определения условий, при которых выполняется определенное действие или подтверждается определенный факт в диалоге.

Проиллюстрируем это на примере, используя приведенный ранее диалог, в котором пользователь выбирает ветку «Назначить время и дату визита в компанию». Логическая структура диалога в этой ветке определяется с помощью событий W_1 , W_2 и W_n , которые соответствуют разным (возможным) вариантам выбора пользователем определенного действия в каждом из возможных миров, где событие такого выбора имело место. Тогда логическое выражение $R(U, A)$ эквивалентно $W_1 \vee W_2 \vee \dots \vee W_n$:

$$R(U, A) = W_1 \vee W_2 \vee \dots \vee W_n,$$

Что значит: «Пользователь записался на прием (R) на 02.02.2024 (W_1), или на 03.02.2024 (W_2), или на иную дату (W_n)»

Теперь, если мы вернемся к исходному логическому выражению: $R(U, A) \wedge T(\text{дата}) \rightarrow P(\text{записаться})$, можно связать его с пропозициональной структурой диалога. Выражение $R(U, A)$ представляет собой пропозицию «Пользователь имеет статус S и намерен записаться на прием N». $T(\text{дата})$ представляет собой пропозицию «Дата записи пользователя на прием соответствует дате T». $P(\text{записаться})$ представляет собой пропозицию «Запись пользователя подтверждена на дату T». В этом случае логическое выражение имеет следующее

значение: «Если пользователь имеет статус S и намерен записаться на прием N и дата записи пользователя на прием соответствует дате T, то запись пользователя подтверждена на эту дату (Wn)».

Приведенная формально-логическая модель коммуникации пользователя и диалогового агента соответствует наиболее простым коммуникативным ситуациям, касающимся информирования пользователя о деятельности компании или помощи в выборе её услуг (товаров). Более интересная модель коммуникации между пользователем и диалоговым агентом возникает в «игровых» формах взаимодействия, когда система выступает в роли собеседника, участвующего в неформальной (дружеской) беседе. В данной коммуникативной ситуации информативная функция уступает место фатической, соответственно повышается значимость тех компонентов коммуникации, которые отвечают за модальность и параметры когниции (например, персонализация диалогового агента, активное ведение диалога, возможность прогнозировать речевые реакции пользователя за счёт выделения ключевых слов или определения тональности высказывания).

Исследование структуры диалога в коммуникации человека и диалогового агента в рамках диалоговой лингвистики должно включать системное исследование промтов (запросов) пользователя как актора коммуникации. Промты, будучи специфическими инструкциями для языковой модели, позволяют выявить зависимость между речевыми запросами и репрезентированными в них коммуникативными интенциями пользователей и качеством диалогового взаимодействия, которое оценивается исходя из критериев его коммуникативной эффективности и успешности.

2.2. Интерфейс автогенерации текстов: лингвистические аспекты анализа промтов

В рамках коммуникативной практики автоматической генерации текста особую роль приобретает стратегия формирования запросов, которые инициируют

процесс создания нейросетью текста, отвечающего параметрам, заданным пользователем.

По нашему мнению, «запрос на генерацию текста, адресованный программным средствам на основе больших языковых моделей, выступает как техническое средство реализации взаимодействия между человеком и машиной и программным интерфейсом системы» [Остапенко, Халина 2025]. Запрос на генерацию текста может быть рассмотрен и в лингвопрагматическом аспекте как сообщение на естественном языке, адресованное интерфейсу программного алгоритма и служащее для решения прикладных задач или развлечения. В связи с этим создание запроса может быть интерпретировано как новая стратегия текстопорождения, ставшая возможной в эпоху развития больших языковых моделей.

История термина «запрос» в русскоязычном научном дискурсе не подвергалась специальной рецепции. Как правило, термин «запрос» употребляют наряду с номинацией «промт» (в иной транслитерации – «промпт») как синонимы. Слово «промт», образованное от англ. *prompt* – «запрос, подсказка», является неологизмом и активно функционирует в дискурсе обучения пользователей основам работы с нейронными сетями. В то же время устоявшееся понимание термина «запрос» (англ. *query* – запрос, вопрос) связана с такой отраслью, как аналитика данных. Термином «запрос» именуют операцию, позволяющую извлекать из баз данных различного рода информацию, отвечающую потребностям пользователя, включая агрегацию данных, применение специализированных поисковых функций, объединение данных и проч. Для реализации этих задач не применяются запросы, созданные на естественном языке, как при работе с нейросетевыми алгоритмами на основе больших языковых моделей, поскольку такие запросы не распознаются программными устройствами. Требуется специализированный структурированный язык запросов (SQL, *Structured Query Language*), который имеет жесткий синтаксис, отличается однозначностью и позволяет производить различные операции с данными.

Формальный логический порядок запросов на языке SQL отличается от свободной структуры промта, который позволяет использовать для взаимодействия с диалоговым агентом естественный язык, а значит, не требует специализированных навыков программирования. Относительная свобода структуры и возможность варьирования элементов промта придают ему более творческий характер и в то же время делают результаты генерации мало предсказуемыми и неоднозначными. Как правило, использование идентичных промтов при обращении к нейросети дает сходные, но не полностью совпадающие результаты, что обусловлено самой природой генерационных процессов, заложенных в нейросетевые алгоритмы.

С лингвистической точки зрения, промт является текстом, обладающим признаками речевого жанра, отличаясь целевой направленностью на результат, целостностью структуры, завершенностью, связью с контекстом и внеязыковой ситуацией общения. Между тем диалогичность промта, присущая речевым жанрам, функционирующим в сфере человеческого общения, специфична: она лишена антропологической природы и предполагает взаимодействие с диалоговым агентом, не обладающим прагматическими установками в коммуникации. Поэтому промт можно рассматривать как разновидность инструктивного жанра, описывающие необходимые параметры текста, который должен быть создан нейросетевым алгоритмом. Вариации этих параметров определяются коммуникативной задачей, стоящей перед пользователем, поэтому могут включать формальные и содержательные аспекты. К первым мы относим параметры длины текста, количества предложений, их визуальное оформление, вторые касаются структурно-композиционных, жанрово-стилевых и языковых особенностей промта.

Основываясь на этих положениях, мы провели исследование на материале, содержащем 200 промтов из открытых коллекций библиотеки запросов в сети Интернет⁶. Результаты позволили описать существующие типы промтов,

⁶ Промты для ChatGPT: нейросети GPT на русском — 163 команды [Электронный ресурс] – Режим доступа: <https://gpt-chatbot.ru/promty-dlya-chatgpt-nejroseti-gpt-na-russkom> (Дата обращения: 28.07.2024).

проанализировать их структурную организацию, композиционно-речевые особенности, а также провести сравнительно-сопоставительный анализ промтов и инструктивных речевых жанров.

Анализ целевого предназначения промтов позволяет говорить о многоаспектной классификации промта, которая включает объект генерации, характер решаемой задачи, тематическое содержание запроса, уровень детализации запроса, степень осведомленности (компетентности) автора запроса в задачах генерации текста. В соответствии с указанными основаниями мы выделяем следующие группы промтов:

1. Промты для генерации различного типа контента: текстов на естественном языке, программного кода, изображений, презентаций, видеоконтента, музыки.

2. Промты, не имеющие исходного образца (создаваемые «с нуля») и направленные на преобразование исходного материала (текстового, графического).

3. Запросы фактического и игрового характера: информационно-поисковый промты, промты для решения образовательных, коммерческих задач, запросы для генерации развлекательного контента.

4. Детализированные, подробные промты и запросы без детализации, с минимальным содержанием требований к процессу и результату генерации.

5. Промты для начинающих пользователей и запросы для опытных («продвинутых») пользователей нейросетей (подробнее о классификации промтов: [Остапенко, Халина 2025]).

Объектом лингвистического исследования может быть любой тип промтов, однако наиболее значимы промты, нацеленные на генерацию текстов на естественном языке. Поскольку задачи текстовой генерации многообразны, разновидности промтов для создания текстов также отличаются разнообразием, включая две обширные группы:

1. Промты для решения задач генерации на основе исходного текста. К этой группе можно отнести автоматический рерайт, редактирование исходного текста, сжатие содержания (суммаризацию), генерацию заголовков к тексту, создание на основе имеющегося текста словарей, списков ключевых слов, плана, а также машинный перевод текста с одного языка на другой.

2. Промты для генерации на основе инструкции пользователя, которые не предполагают преобразование исходного текста, а направлены на создание нового текста, удовлетворяющего заданным параметрам.

Промты второй группы можно считать специфическим типом инструкции для работы с нейросетевым алгоритмом для достижения цели генерации. Основу этой инструкции составляет последовательность компонентов, равнозначных системе формальных действий, направленных на решение задач генерации.

Нами установлено, что формализация промта достигается за счет создания его типовой структуры, включающей заголовок и тело промта. Заголовок промта указывает на его тему и характер решаемой задачи (например, «Генерация заголовков для статьи», «Составление расписания дня», «Написание рецензии» и др.) Запрос формулируется в теле промта и оформляется в виде побудительного предложения или последовательности таких предложений [Остапенко, Халина 2025]. Часто в теле промта используются глаголы «напиши», «перепиши», «составь», «придумай» в сочетании с прямым дополнением, обозначающим объект генерации: «перепиши текст», «напиши статью», «исправь ошибки», «составь план», «придумай 5 вариантов заголовков».

Некоторые исследователи [Полоян, Приходько 2024] считают, что существует трехчастная структура промта для генерации изображений, включающая сюжетное, художественное и техническое описание будущего изображения. В промтах для генерации текста сюжетная часть факультативна и может применяться в тех случаях, когда необходимо создать текст повествовательного характера, пошаговое руководство или план. Художественная часть эквивалентна указанию жанрово-стилевых и композиционно-речевых особенностей будущего текста. Техническая составляющая позволяет задать

формальные характеристики генерируемого текста (длина, количество абзацев, особенности визуального оформления).

Нам представляется методологически более корректным подход к структуре промта, основанный на применении спецификации как перечня «коммуникативно-речевых параметров, которым должен соответствовать генерируемый текст» [Остапенко, Халина 2025].

Анализ общедоступных промтов позволяет выделить базовые элементы, составляющие его спецификацию:

1. Прагматический компонент промта определяется адресной направленностью диалогового агента, который выступает в определённой коммуникативной роли как «тренер», «консультант», «преподаватель», «психолог», «журналист» и так далее.

2. Содержание промта зависит от тематики запроса, поэтому является наиболее вариативной частью спецификации и представлено в виде задачи генерации. Часто по форме это побудительное предложение или описание желательных для пользователя действий нейросетевого алгоритма.

3. Формальный компонент промта связан с дополнительными параметрами запроса, определяющими стилевые, жанровые, эмоционально-экспрессивные, речевые характеристики генерируемого текста, а также ограничения или запреты, накладываемые пользователем на генерационные процессы (например, ограничение по объёму) (Таблица 2).

Таблица 2. Спецификация промта для генерации текста⁷

№ п\п	Компонент промта	Примеры промтов
1.	Прагматический компонент	«Выступи в роли учителя...», «Выступи в роли психолога...», «Выступи в роли мотивационного тренера...»
2.	Содержание промта (задача	«Ваша задача сгенерировать сложный

⁷ Источник примеров: Библиотека промтов – Промты для ChatGPT на русском [Электронный ресурс] – Режим доступа: <https://prompt-chatgpt.ru/biblioteka-promptov/> (Дата обращения: 24.07.2024).

№ п\п	Компонент промта	Примеры промтов
	генерации)	пароль с помощью этих форм ввода и предоставить его мне», «Я буду вводить команды, а вы будете отвечать тем, что должна показать консоль javascript», «Я предоставляю вам подробную информацию об организации, нуждающейся в помощи в проектировании или обновлении веб-сайта, а ваша роль заключается в том, чтобы предложить наиболее подходящий интерфейс и функции, которые могут улучшить взаимодействие с пользователем, а также достичь бизнес-целей компании»
3.	Формальный компонент:	
3.1.	жанр	«Вы можете выбрать любой жанр, например, фэнтези, романтику, историческую фантастику и так далее, но цель – написать что-то с выдающейся сюжетной линией, увлекательными персонажами и неожиданной кульминацией»
3.2.	речевое оформление	«Предложения должны быть грамматически правильными и подходить для студентов со средним уровнем владения английским языком»
3.3.	эмоциональная окраска	«Пишите на любую тему, но убедитесь, что ваши слова передают чувство, которое вы пытаетесь выразить, красиво и в то же время содержательно»
3.4	исключения и запреты	«Ваши рабочие листы не должны содержать никаких объяснений или

№ п\п	Компонент промта	Примеры промтов
		дополнительных инструкций, только список предложений и варианты слов»

Мы полагаем, что языковая организация промта определяется сферой его функционирования и предназначением. Речевое оформление промта характеризуется следующими особенностями:

1. Лексический состав промтов, как правило, однотипен и включает преимущественно конкретную лексику, лишенную образности.

2. Синтаксическая организация промта отличается формальной структурой. Наиболее частотны простые побудительные предложения с глаголом повелительного наклонения в качестве сказуемого. Формальный синтаксис промта подразумевает и однотипное построение предложений, в которых доминирует прямой порядок слов.

3. Сложные предложения в промтах используются редко, в основном для конкретизации задачи генерации. В этом случае используются сложноподчиненные конструкции с придаточными цели («Перепиши данный текст, чтобы улучшить его читаемость...») или с придаточными определительными («Напиши текст, который будет соответствовать заданному стилю и тону, сохраняя основное содержание...»).

4. Для синтаксической структуры промта также характерно «нанизывание» простых предложений в составе сложного для описания задачи генерации («Перепиши данный текст, сохрани основную идею, добавь новые факты, сделай его более структурированным», «Составь инструкцию по использованию нового программного обеспечения, опиши основные функции, добавь пошаговые рекомендации для пользователя») (подробнее об этом: [Остапенко, Халина 2025]).

Специфика речевой организации промта заключается также в экспликации в его языковом оформлении концепта «речевое (речемыслительное) действие», поскольку промт ориентирован на выполнение ряда задач, соотносимых с теми,

которые решаются в рамках когнитивной деятельности. Языковые средства, описывающие данную когнитивную деятельность, употребляются метафорически, приписывая нейросети специфически человеческие ментальные способности. К наиболее часто встречающимся маркерам «гуманизации» промта можно отнести:

глаголы с речемыслительной семантикой, употребляющиеся в отношении человека: «напиши», «составь», «придумай», «вы будете отвечать», «вы посоветуете», «вы будете комментировать» и так далее;

устойчивые сочетания, приписывающие программе человеческие модели поведения: «выступи в роли...», «вы также должны обязательно включить ... личные анекдоты или опыт», «вы также должны использовать свои риторические знания и опыт в области эффективной техники письма», «ваша цель – помочь людям выйти из дискуссии с более глубокими знаниями и пониманием рассматриваемой темы», «важно подчеркнуть, какие чувства вызвал у вас фильм. Что действительно вызвало у вас отклик. Вы также можете критически отозваться о фильме», «вы должны использовать свои знания в области психологии животных и методов модификации поведения для создания эффективного плана»;

прямое отождествление с человеком с помощью вербальных маркеров «человек», «кто-то»: «Нужен человек с опытом работы с автомобилями в отношении поиска и устранения неисправностей», «Мне нужен человек, который мог бы предложить вкусные рецепты, включающие полезные продукты», «Нужен кто-то достаточно опытный, чтобы различать различные виды чая на основе вкусового профиля»;

прилагательные с оценочной семантикой для характеристики результата действий, приписываемых алгоритму: «ваша задача – объяснить их *простыми и понятными* словами», «вы напишете статью, в которой дадите *проницательный* комментарий по данной теме», «Вы будете предлагать мне *веселые, уникальные* идеи для занятий и хобби, которые можно делать где угодно» [подробнее: Остапенко, Халина 2025].

Лежащая в основе данных вербальных репрезентаций метафора «человек – машина», позволяет говорить о наличии связи между генерационными

процессами и речевыми операциями, которые имеют антропологическую природу. Объединяясь в аспекте порождения высказывания на естественном языке, эти процессы демонстрируют значимость лингвокогнитивной организации знаний в структуре запроса для генерации и актуализируют проблему выбора стратегий коммуникативного поведения субъекта при взаимодействии с программным алгоритмом в русле промт-инжиниринга.

2.3. Промт-инжиниринг как стратегия текстопорождения

Промт-инжиниринг (проектирование промтов) представляет собой методику разработки и оптимизации текстовых запросов, которые обеспечивают процесс взаимодействия пользователя и нейросетевого алгоритма в ходе генерации⁸. Как отмечают разработчики, промт-инжиниринг «охватывает широкий спектр навыков и техник, полезных для взаимодействия с LLM и создания новых решений на их основе»⁹, следовательно, имеет не только прикладной, но и исследовательский характер.

Промт-инжиниринг может быть рассмотрен как совокупность техник улучшения формально-содержательной структуры запроса на естественном языке, влияющих на коммуникативную эффективность и качество сгенерированного текста. Следовательно, процессы, связанные с проектированием промтов как феноменов, использующих ресурсы естественного языка для решения широкого спектра задач взаимодействия пользователя и нейросетевого алгоритма, тесно связаны с новыми механизмами производства текста с участием компьютерных алгоритмов.

Инжиниринг промтов опирается на формальные модели текстовых структур, представленные в виде шаблонов. «Шаблон промта — это незаконченный промт с одной или несколькими переменными, которые

⁸ Руководство по проектированию промтов – Yandex Cloud [Электронный ресурс] – Режим доступа: https://yandex.cloud/ru/docs/foundation-models/gpt-prompting-guide/about?utm_referrer=https%3A%2F%2Fyandex.ru%2F (Дата обращения: 24.07.2024).

⁹ Там же.

заменяются на какую-то информацию для создания отдельного экземпляра промта»¹⁰. Варьирование переменных в шаблоне промта определяется входными и выходными параметрами запроса, включающими набор исходных данных для генерации (например, наличие текста-образца) и предполагаемый результат генерации (тематическая направленность, адресация искусственного текста, его стиль, структура).

Процесс конструирования промтов включает:

1. Определение целей и задач генерации, а также соответствующих им лингвистических и экстралингвистических параметров, характеризующих переменные для конструирования запроса.
2. Выбор техник, которые оптимизируют работу модели, улучшая результат генерации. К ним можно отнести конструирование подсказок, в том числе содержащих несколько примеров или образцов для генерации, определение последовательности шагов, имитирующих цепочку рассуждения (chain-of-thought (CoT) prompting)¹¹, использование результатов критической оценки алгоритмом собственного решения для усовершенствования структуры и содержания запроса (self-refine)¹².
3. Генерации текстов на основе различных промтов для проверки их релевантности решаемой коммуникативной задаче.
4. Оценка полученных результатов генерации на основе соотнесения с прагматическими установками субъекта для выбора оптимальных параметров промта, влияющих на качественные характеристики генерируемого текста.

¹⁰ Там же.

¹¹ Wei J., Wang X., Schuurman, D., Bosma M., Ichter B., Xia F., Chi Ed H., Le Quoc V., Zhou D. (October 31, 2022). Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *Advances in Neural Information Processing Systems (NeurIPS 2022)*. Vol. 35. arXiv:2201.11903.

¹² Madaan A., Tandon N., Gupta P., Hallinan S., Gao L., Wiegreffe S., Alon U., Dziri N., Prabhume S., Yang Y., Gupta Sh., Prasad M. B., Hermann K., Welleck S., Yazdanbakhsh A. (March 1, 2023). Self-Refine: Iterative Refinement with Self-Feedback. arXiv:2303.17651.

5. Использование полученного промта в серии генераций для оценки его пригодности при создании текстов, которые отвечают различным параметрам, включающим соблюдение речевых норм естественной коммуникации, содержательной и фактической точности, достоверности, этических требований к результатам генерации.
6. Составление на основе наиболее эффективных для генерации промтов коллекций или корпусов, которые могут быть использованы широкой аудиторией или персональным пользователям в целях генерации текстового контента.

Описанный процесс конструирования промтов характерен как для разработчиков, тестирующих новые решения, так и для обычных пользователей, которые могут улучшать работу алгоритма, совершенствуя механизмы промт-инжиниринга при решении конкретной задачи генерации.

Использование шаблонов промтов обеспечивает релевантность и предсказуемость результатов генерации текстов, поскольку позволяет формализовать структуру запроса, минимизируя при этом риск возможных недочетов или ошибок, которые могут быть допущены пользователем. Успешность формализации промта с помощью методов формально-логического анализа языка продемонстрирована в исследовании [Остапенко, Халина 2025]. Шаблон промта выступает как «логическое выражение с использованием переменных, кванторов и предикатов. Формализация промта в теории предикатов включает рассмотрение всех его компонентов как элементов логической структуры высказывания, включающей переменные и константы, кванторы существования, предикаты (в качестве атрибутов – свойств объекта), логические выражения.

В качестве примера формализации может быть рассмотрена логическая структура нескольких промтов для решения типовых задач генерации текстов на основе общедоступных коллекций запросов от «Яндекса»¹³ (Таблица 3).

Таблица 3. Формально-логическая структура типов промтов

Задача генерации	Переменные и константы	Кванторы существования	Предикаты	Логическое выражение
Краткий пересказ (суммаризация) текста	T (Text / Текст) S (Summary / Суммаризация)	$\exists T$ – существует текст $\exists S$ – существует его краткий пересказ	Summarizes (T, S) – предикат «суммаризация S текста T»	$\exists T \exists S (\text{Summarizes}(T, S))$
Извлечение информации	T (Text / Текст) I (Information / Информация) C (Context / Контекст)	$\exists T$ – существует текст $\exists I$ – существует информация $\exists C$ – существует контекст значимости информации	Extracts(T, I) – предикат «информация I извлекается из текста T». BasedOn(T, C) – предикат «извлечение информации I основано на контексте C»	$\exists T \exists I \exists C (\text{Extracts}(T, I) \wedge \text{BasedOn}(T, C))$
Ответ на вопрос	C (Context / контекст) Q (Question / вопрос) A (Answer / ответ)	$\exists C$ – существует контекст $\exists Q$ – существует вопрос $\exists A$ – существует ответ	Answers(Q, A) – предикат «ответ A на вопрос Q основан на контексте C»	$\exists C \exists Q \exists A (\text{Answers}(Q, A) \wedge \text{BasedOn}(C))$
Классификация текста	T (Text / текст) P (P / Параметр классификации)	$\exists T$ – существует текст $\exists P$ – существует параметр, по	Classifies(T, P) – предикат «текст T классифицируется по параметру P»	$\exists T \exists P (\text{Classifies}(T, P))$

¹³ Источники примеров: Руководство по проектированию промтов – Yandex Cloud [Электронный ресурс] – Режим доступа: https://yandex.cloud/ru/docs/foundation-models/gpt-prompting-guide/about?utm_referrer=https%3A%2F%2Fyandex.ru%2F (Дата обращения: 24.07.2024).

Задача генерации	Переменные и константы	Кванторы существования	Предикаты	Логическое выражение
		которому текст может быть классифицирован		
Поддержание разговора	U (User / пользователь) A (Assistant / ассистент) R (Replica / Реплика)	$\exists U$ – существует пользователь $\exists A$ – существует ассистент $\exists R$ – существует реплика	Responds(A, R) – предикат «ассистент A отвечает на реплику R пользователя U»	$\exists U \exists A \exists R (\text{Responds}(A, R))$

Формализация языка промтов, позволяющая создавать шаблоны запросов, сталкивается с проблемой двусмысленности и неопределённости естественного языка, который используется для взаимодействия пользователя и программного алгоритма, поэтому на первый план выходят механизмы лингвистического конструирования запроса, который релевантен решаемой задаче генерации и позволяет получать приемлемые результаты. Эта проблема, как мы считаем, может быть решена с помощью разработки методов лингвистической оптимизации промтов для задач генерации текста.

Нами предложено понятие методологии лингвистической оптимизации промта, под которой понимается «системный подход к конструированию запроса (промта), учитывающий совокупность лингвистических переменных, необходимых для повышения коммуникативного качества и точности ответов языковой модели на основе искусственного интеллекта» [Остапенко, Халина 2025: 46].

Экспериментальный подход к разработке методологии лингвистической оптимизации промта на естественном языке представлен в нами в исследовании [Остапенко, Халина 2025] и основан на проверке релевантности запросов, сконструированных с разным набором параметров, сгенерированному тексту в

аспекте его коммуникативных качеств и формально-содержательных характеристик с использованием количественной и качественной оценки.

Качественный анализ предполагает соответствие сгенерированного текста критериям соответствия тематике запроса, фактологической точности, связности и цельности, грамматической корректности. Оценка по каждому критерию производилась вручную по градуальной шкале, позволяющей определить степень выраженности критерия в анализируемом тексте.

Количественные параметры оценки включали набор метрик, позволяющих оценить индекс лексического разнообразия текстов, их лексическую плотность, а также показатели текстовой информационной энтропии как меру семантического разнообразия текстов. Они были вычислены автоматически с помощью программного кода на языке программирования Python. Использование количественных метрик при сопоставительном анализе содержательного и лексического разнообразия сгенерированных на основе различных промтов текстов позволяет подтвердить и дополнить качественные параметры оценки. Более подробное описание используемых количественных метрик оценки приведено в Таблице 4.

Таблица 4. Количественные метрики оценки сгенерированных текстов

Метрика	Описание
MTLD (The Measure of Textual Lexical Diversity) Индекс лексического разнообразия	$MTLD = \frac{M}{T}$, где U – уникальные токены, T – общее количество токенов [McCarthy, 2010]. Метрика позволяет вычислить индекс лексического разнообразия с учетом средней длины текстов.
Индекс лексического разнообразия	$TTR = \frac{U}{T}$, где U – количество уникальных слов, T – общее количество слов [Ure, 1969]. Метрика позволяет оценить индекс лексического разнообразия без учета длины текста.
Лексическая плотность (отношение полных слов к служебной лексике)	$LD = \frac{N_{content}}{N_{total}}$, где $N_{content}$ – количество полных слов, N_{total} – общее количество слов. Метрика позволяет количественно оценить степень информативности текста.
Показатель текстовой информационной энтропии	$H(x) = -\sum_{i=1}^n p_i \log_2 p_i$, где p_i – вероятность i -го слова в тексте.

Метрика	Описание
	Метрика позволяет вычислить суммарные показатели вероятности появления слова в тексте для определения информативных «потерь» за счет часто встречаемых слов.

Варьирование шаблона промтов, использованных для генерации текстов, основано на сочетании лингвистических переменных, касающихся синтаксической структуры, семантики и прагматики запроса (Таблица 5).

Таблица 5. Типы переменных, используемых для конструирования промта

Группа лингвистических переменных	Параметры оптимизации промта
Синтаксические переменные	Тип синтаксической конструкции: вопросительная / императивная конструкция
Семантические переменные	Ключевые слова текста
Прагматические переменные	Жанрово-стилевые параметры текста, сфера его функционирования
Экстралингвистические факторы	Длина запроса, указание на желаемую структуру генерируемого текста

В ходе экспериментального исследования проведена серия из 15 генераций текстов с помощью модели LLM ChatGPT на основе вариантов запросов, различающихся степенью детализации при выборе указанных параметров. Запросы были сконструированы вручную на основе случайно выбранных тем из массива общедоступных направлений (образовательная тематика, медиа) с учетом вариативности функциональной структуры текста (описательный, аналитический или диалогический текст). Полученные в ходе эксперимента варианты текстов представлены в Приложении 1.

В соответствии с выбранной методологией экспериментального исследования они были проанализированы с учетом выбранных качественных и количественных метрик оценки для определения структуры промтов, позволяющих производить тексты приемлемого коммуникативного качества.

Качественный анализ сгенерированных текстов вне зависимости от структуры запроса продемонстрировал их тематическую релевантность запросу, фактическую точность и достоверность, смысловую связность и цельность, а также грамматическую корректность (имеющиеся аграмматизмы носят единичный характер и находятся в пределах погрешности, которая может быть допущена человеком при составлении текста). Количественная оценка сгенерированных текстов показывает, что критерии информативности и лексического разнообразия чувствительны к настройке параметров запроса. Так, более детальные запросы, которые включают все переменные в шаблоне промта, увеличивают длину генерируемого текста и индекс лексической плотности за счет увеличения общего количества полнозначных слов, однако снижают показатели информативности, лексического и семантического разнообразия, как показано в Таблице 6.

Таблица 6. Среднее значение метрик сгенерированных текстов для промтов различной степени детализации

Степень детализации промта	Группы текстов	MLTD	Лексическая плотность	Энтропия текста	Индекс лексического разнообразия
Низкая	Группа №1	120,09	0,29	5,14	0,88
Средняя	Группа №2	121,89	0,37	5,39	0,83
Высокая	Группа №3	101,55	0,61	5,73	0,70

Анализ полученных данных свидетельствует о том, что при избыточной параметризации промта генерируемый текст теряет информационную насыщенность за счет увеличения доли ключевых слов из запроса, а также лексических повторов в заключительной части текста, дублирующих его основное содержание.

Лингвистическая оптимизация промта, в отличие от формализации его структуры, основана на экспериментальном варьировании параметров запроса, которое основано на выборе оптимального сочетания требований, предъявляемых

к искусственному тексту. Параметры, направленные на контекстуализацию запроса, структурирование текста, определение его жанровых и стилевых характеристик, должны применяться с учетом объема генерируемого текста, его содержательной глубины и коммуникативной направленности. Рядовые задачи генерации эффективно решаются с помощью формализованных запросов (шаблонов промтов), обеспечивая оптимальные коммуникативные характеристики генерируемых текстов. Конструирование промтов приобретает значимость в ситуациях, когда необходимо повысить метрики текста, отражающие уровень его лексико-семантического разнообразия и текстовой энтропии, то есть в ситуациях лингвокогнитивного структурирования знания для решения коммуникативной задачи. Эти процессы связаны с механизмами когнитивной инженерии и шире – лингвистической инженерии [Халина, Злобина, Авдеева 2014].

В этом аспекте промт-инжиниринг выступает как механизм репрезентации знаний субъекта о свойствах генерируемого текста, который одновременно является механизмом обучения и корректировки для системы искусственного интеллекта. Промт в этом случае выступает как метатекст, позволяющий определить коммуникативные параметры модели производства текста и прогнозировать его содержание на основе имеющегося речевого опыта. Следовательно, качество промта, как и эффективность процесса генерации, зависит не только от технических характеристик модели, но степени осведомленности пользователя о коммуникативной ситуации, в рамках которой формируется задача генерации текста.

Таким образом, промт-инжиниринг как стратегия конструирования промтов является областью междисциплинарных исследований на стыке решения инженерных и лингвокогнитивных задач, тесно смыкаясь с проблемой повышения коммуникативных качеств сгенерированного текста.

2.4. Информационная модель содержания сообщения в контексте генерации текста

Информационное содержание сообщения сгенерированного сообщения заслуживает особого внимания, так как изменение стратегий текстопорождения в связи с появлением больших языковых моделей, повлекло за собой и изменение механизмов организации и репрезентации информации в тексте. Отталкиваясь от сложившегося в лингвистике представлений о моделировании информационного содержания сообщения как метода изучения речевой деятельности человека, следует отметить её трансформацию с появлением гибридных форм коммуникации, которые охватывают процессы речевого взаимодействия человека и диалоговых агентов – автономных систем, функционирующих на основе генеративных алгоритмов. Указанные информационные модели сообщения представлены в основном в вопросно-ответной форме и обладают интерактивной природой, позволяющей вычленять компоненты модели, относящейся к действию субъекта (пользователя) и алгоритма. Лингвистическим механизмом, обеспечивающим возможность такого взаимодействия, выступает транскодинг как система, позволяющая производить кодировку-декодировку сообщений в системе кодов, доступных машинному агенту и антропологическому субъекту.

Компоненты информационной модели сообщения в контексте генерации включают входные данные и этапы их преобразования машинным алгоритмом, позволяющие получить результирующий текст на естественном языке, который доступен для восприятия и интерпретации наравне с антропологическими текстами. Промежуточные процессы преобразования входных текстовых данных скрыты от пользователя, поэтому на эмпирическом субъекту доступны только два компонента такой модели: входные данные (запрос) – результирующее сообщение (ответ), которое обладает определённой семантикой и может быть интерпретировано в контексте запроса как отвечающее его условиям или требующее корректировки. В то же время нейросетевому алгоритму не доступны те компоненты информационной модели, которые очевидны для пользователя и

составляют его прагматический потенциал в контексте решения той или иной задачи генерации.

Информационная модель содержания сгенерированных сообщений на естественном языке опосредована как задачей генерации, так и коммуникативным предназначением создаваемого текста, который будет функционировать в определённом типе дискурса. Поскольку сгенерированные тексты преимущественно используются в среде интернет-коммуникации, компоненты их коммуникативной модели имеют особенности, обусловленные рождением и функционированием текста цифровой природы. К наиболее значимым компонентам коммуникативной модели сгенерированного текста можно отнести коммуникативный код сообщения, модус сообщения, актуализирующий канал передачи информационного содержания, а также коммуникативную ситуацию их создания, включая исходные данные, представленные в запросе для генерации, а также функционал автономного диалогового агента, определяющий его роль во взаимодействии с пользователем. В целом указанные компоненты коммуникативной модели соотносятся с традиционной структурой коммуникативной модели сообщений в естественной коммуникации, однако имеют свои особенности:

1. Коммуникативный код, используемый для передачи сгенерированного сообщения, неоднороден и основывается на принципе перекодировки в цепочке «код естественного языка – машиночитаемый код (обобщенное векторное представление текста) – код естественного языка».

2. Модус сгенерированного сообщения опирается на рецепторные возможности субъекта, задействуя вербальные и невербальные каналы восприятия информации. К последнему можно отнести зрительный и слуховой каналы восприятия сообщений, которые приобретают особую значимость в системах генерации поликодовых текстов (например, презентаций, инфографики, фото- и видеоизображений, сопровождающих текстовой подписью).

3. Коммуникативная ситуация производства сгенерированного сообщения не ограничивается простым взаимодействием пользователей диалогового агента по

принципу «стимул (запрос) – реакция (ответ)». Она приобретает динамический и интерактивный характер, поскольку современные диалоговые агенты обладают адаптивным функционалом, позволяющим не просто отвечать на запрос пользователя, но и учитывать его контекст и пользовательскую корректировку запрашиваемой информации.

В результате возникает особый тип коммуникативной ситуации, результатом которой выступает сообщение, обладающее эргодичностью, то есть ориентированное на принцип сотворчества с субъектом при его создании. При этом правила создания сгенерированного текста изначально заданы алгоритмической структуры, однако вариант их комбинации или реализации, приводящий к текстогенерации, выбирается субъектом всякий раз, когда он определяет их в запросе. За счет заложенной в алгоритмы вариативности при производстве вычислений результат генерации на один и тот же запрос может быть отличаться, что приводит к появлению такого текстового свойства, как мутабельность (изменчивость) под влиянием взаимодействия с субъектом. Этим свойством, в частности, обусловлена и свобода субъекта в выборе стратегий освоения информационного содержания текста, который определяется индивидуальной логикой взаимодействия человека и алгоритма в рамках заранее заданной и многовариантной системы исчисления результата генерации.

Анализ эмпирических данных, отражающих коммуникативно-языковые особенности текстовых сообщений в процессе их генерации различными диалоговыми агентами, позволяет выстраивать их типологию на основе способов представления информационного содержания. В качестве оснований для типологии можно рассмотреть различные типы модальностей и модусы восприятия информации. В соответствии с принятыми критериями можно выделить:

1. Сообщения на основе комбинирования модальностей, включающие примеры сгенерированных текстов, в которых визуальный компонент (например, изображения или графики) дополняется текстовой информацией. В диалоговых системах примером такого сообщения является запрос на генерацию текстовых

подписей к изображениям (рис. 6), генерация текста, замещающего изображение (рис. 7), а также создание поликодовых сообщений разных жанров (презентации, видеоролики) (рис. 8).

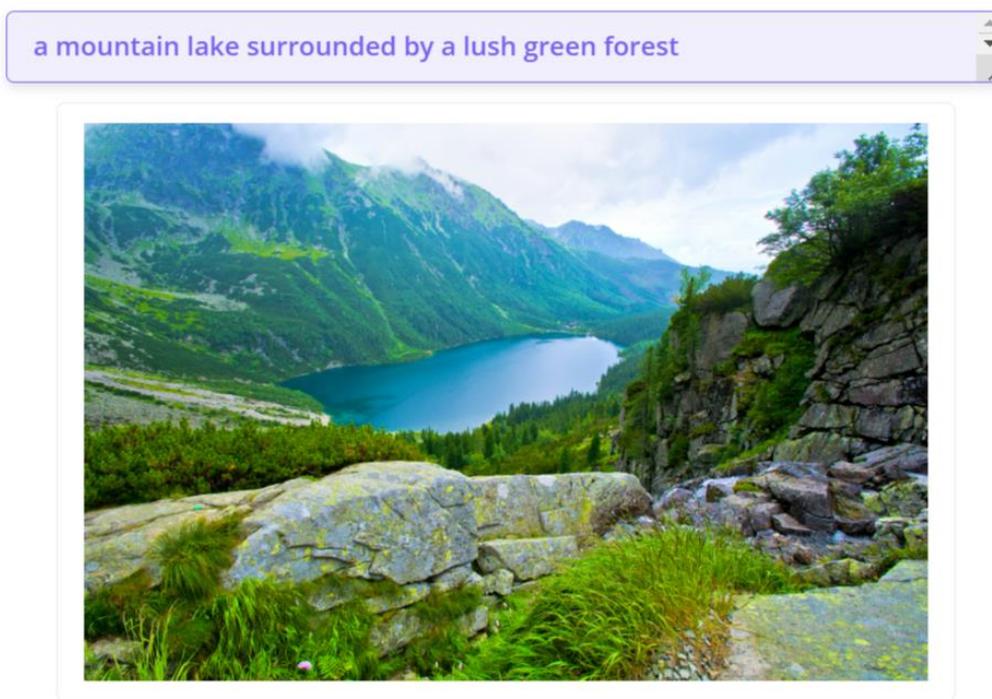


Рисунок 6. Сгенерированная подпись к изображению

Источник: ИИ-сервис i2img.com

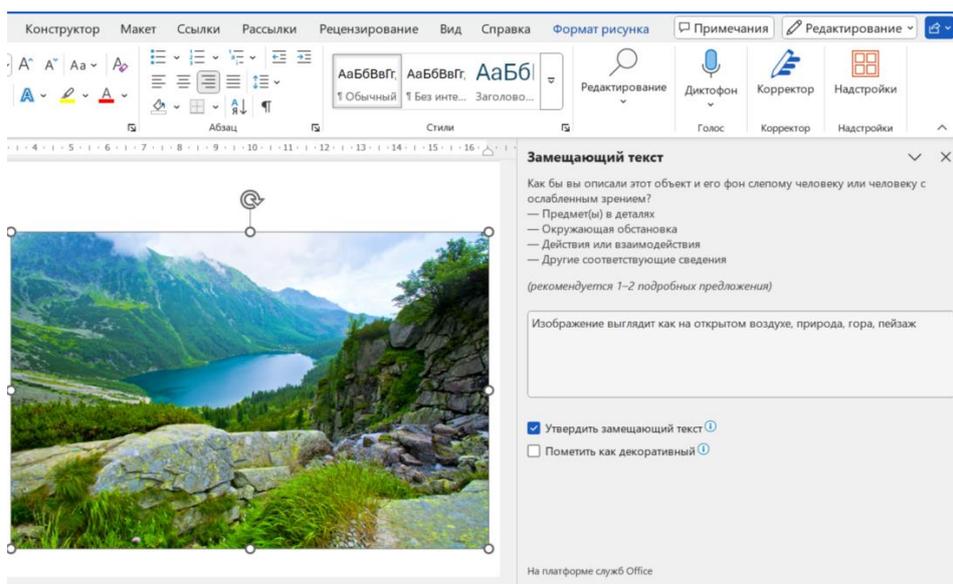


Рисунок 7. Сгенерированный замещающий текст к изображению

Источник: Microsoft Office с ИИ-ассистентом

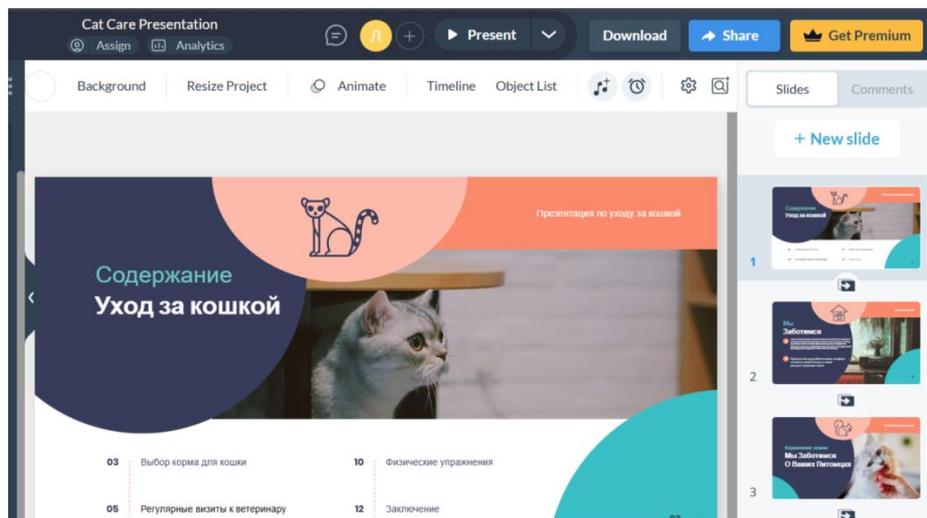


Рисунок 8. Сгенерированная презентация

Источник: ИИ-генератор visme.com

Подобные информационные сообщения, как видно из анализа представленных примеров, имеют гетерогенную кодовую природу и сочетают несколько модальностей – визуальную (графическую) и вербальную. Такие тексты в лингвистике известны как креолизованные и (или) поликодовые. При этом содержание вербальной составляющей такого текста формируется на основе запроса пользователя. Пример с генерацией презентации на основе запроса пользователя (промт: «Создать слайды по тематике ухода за кошкой») демонстрирует единство визуальной и текстовой информации, формирующей целостное информационное сообщение, отражающее разные аспекты темы, отличающееся логичностью, последовательностью развертывания информационного содержания, соответствием визуального и вербального компонентов сообщения.

Каждый из указанных способов репрезентации сгенерированного сообщения на основе сочетания различных модальностей предполагает мутабельность (изменчивость) конечного сообщения за счет механизма корректировки результата со стороны пользователя. Поэтому выбор содержания сообщения, его языковой репрезентации и визуального стиля оформления

определяется пользователем, иницируя процесс сотворчества с нейросетевым агентом.

2. Сообщения на основе синхронизации различных модусов восприятия включают сгенерированные тексты, обеспечивающие скоординированную передачу информационного содержания через разные каналы восприятия. Примером таких сообщений выступают примеры диалогового взаимодействия с «умными помощниками» (ИИ-ассистентами), где голосовой ответ агента сочетается с текстовой и визуальной составляющей (рис. 9).

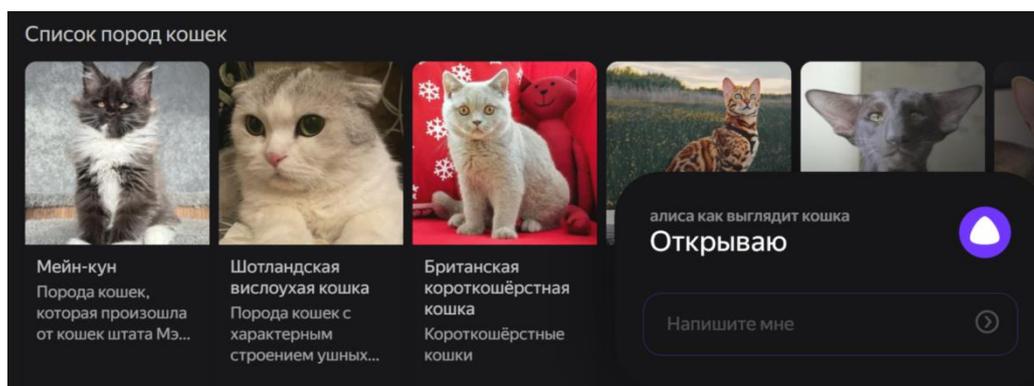
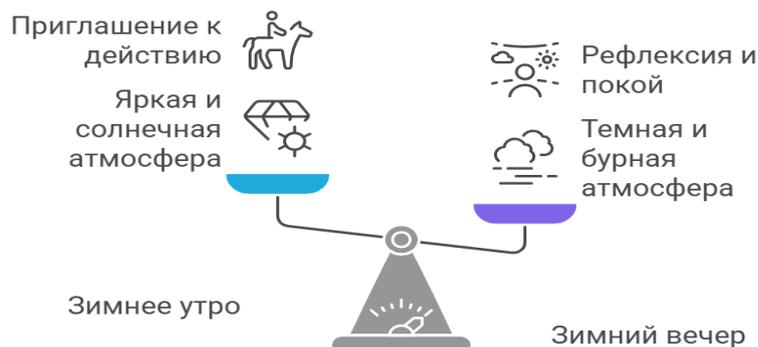


Рисунок 9. Поисковая выдача голосового помощника на запрос информации пользователя

Источник: голосовой помощник «Алиса» (браузер «Яндекс»)

Анализ сгенерированных сообщений такого рода показывает тенденцию к использованию различных модусов восприятия информации, которые сочетают визуальные вербальные и невербальные и слуховые каналы передачи содержания сообщений. Характерной коммуникативной особенностью данного типа сообщений является их соответствие кибертексту, который не только предполагает выбор стратегии нелинейного продвижения в структуре сообщения, но и подразумевает активность субъекта в плане интеграции различных компонентов содержания для получения целостного представления на основе агрегированной информации.

3. Сообщения на основе перекодировки исходного текста с целью суммаризации и репрезентации его информационного содержания в виде схем, инфографики, ментальных карт, «облака слов» и проч. (рис. 10–11).



Сравнение зимних утренних и вечерних настроений в поэзии Пушкина.

Рисунок 10. Перекодировка вербальной информации в графическую (на примере стихотворения А. С. Пушкина «Зимнее утро»)

Источник: ИИ-сервис parin.ai

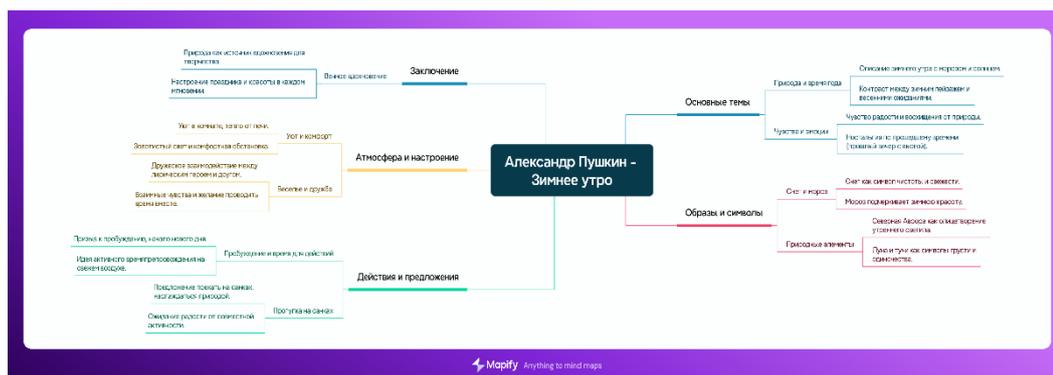


Рисунок 11. Ментальная карта, сгенерированная на основе стихотворения А. С. Пушкина «Зимнее утро»

Источник: Суммаризатор ментальных карт на базе ИИ marify.so

Специфика данных сообщений состоит в том, что в процессе генерации производится структурно-тематическое моделирование исходного текста, позволяющее вычленять ключевые смысловые паттерны с последующей их визуализацией для сокращения текста и представления его в виде схемы, в которой пользователю предлагается определенный формат подачи информации и способы интерпретации смысловых связей исходного текста.

4. Сообщения с динамической структурой, адаптирующиеся к контексту и цели коммуникации и изменяющиеся в зависимости от взаимодействия пользователя с системой. В качестве наиболее репрезентативного примера можно привести сгенерированные тексты поисковой выдачи, адаптирующиеся под параметры пользовательского запроса и формирующиеся мгновенно как ответ поисковой выдачи на основе механизма агрегации информации по теме (рис. 12–13).

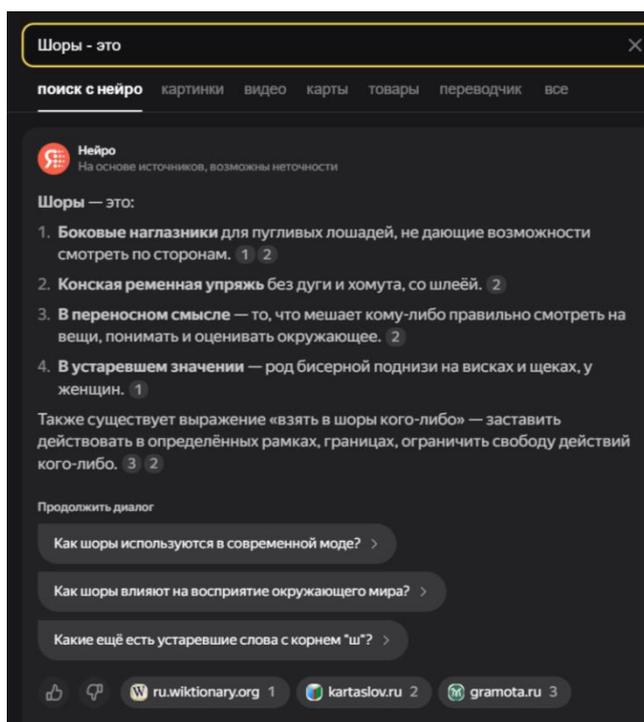


Рисунок 12. Результат нейропоиска по запросу на основе суммаризации источников по теме

Источник: функция «Нейропоиск» (браузер «Яндекс»)

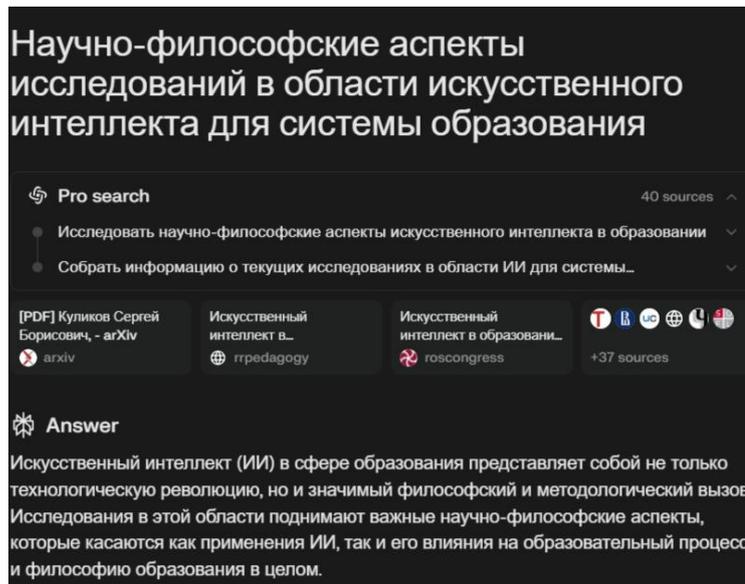


Рисунок 13. Ответ на запрос с функцией агрегации источников по теме

Источник: поисковая система Perplexity AI

Последний тип сгенерированных сообщений отличается непостоянством структуры и содержания, которые трансформируются под влиянием исходных источников суммаризации. Гипертекстовая структура позволяет переходить по гиперссылкам к развёртыванию содержательно-смысловых компонентов сообщения и задает нелинейный способ чтения. Промты для продолжения диалога, которые встраиваются в текстовую структуру в качестве веток сценария возможного взаимодействия с поисковой системой, актуализируют стратегию выбора пользователем информационно-содержательных аспектов генерируемого текста.

Различные типы сгенерированных сообщений в процессе взаимодействия субъекта и диалогового агента обладают общими признаками, которые можно интерпретировать как принципы информационного моделирования данных сообщений. К ним можно отнести:

1. Мультиmodalность, связанную с использованием негомогенных кодов для формирования сообщения, сочетающего вербальный код и визуально-графические и (или) аудиальные компоненты.
2. Редундантность как признак информационной избыточности сообщений, в которых одно и то же информационное содержание передается с помощью

различных кодов и каналов восприятия: письменная текстовая информация дублируется звуковой, визуально-графическая – текстовой.

3. Симультанность, то есть мгновенность формирования текста в ответ на запрос пользователя. Тексты сгенерированных сообщений не существуют до реализации нейросетевым алгоритмом операций по исполнению пользовательского запроса (вернее, они существуют потенциально как правила исчисления возможных вариантов ответа на запрос, заданные структурой алгоритма).

3. Интерактивность как свойство стимулирования активности субъекта при восприятии и интерпретации сгенерированного текста. Структурно-содержательный динамизм сгенерированного сообщения отражает адаптивную природу текстов нового типа, которые формируются и структурируются в соответствии с информационными и коммуникативными потребностями пользователя.

По нашему мнению, обобщая указанные принципы, можно перейти к моделированию информационного содержания сгенерированных сообщений. Компонентами данной модели выступают:

1. Семантический компонент, то есть смысловое содержание сообщения, обеспечивающее передачу информации, включая генеральную тему или совокупность микротем в составе общей темы.

2. Структурно-логический компонент, к которому относятся семантико-синтаксические связи элементов текста сообщения и его логико-композиционная структура.

3. Языковые атрибуты, включая стилевые особенности, маркеры языковой экспрессии, характер используемых языковых средств (например, ключевых слов).

4. Форма репрезентации сообщения (например, сгенерированное сообщение может выступать в виде текста на естественном языке, креолизованного текста, кибертекста и проч.).

5. Функционально-прагматический компонент, который подразумевает способ понимания и восприятия сообщения пользователем.

6. Метаданные, включающие сведения об используемых источниках генерации и параметрах запроса.

Условно модель может быть представлена в виде схемы (рис. 14).



Рисунок 14. Модель информационного содержания сгенерированного сообщения

Процесс генерации сообщения диалоговой системой представляет собой динамическое взаимодействие указанных компонентов, основанное на перекодировке текстовой информации в машиночитаемый формат с последующей репрезентацией результатов выдачи в форме, доступной для восприятия и интерпретации субъектом. Источником сообщения выступают метаданные, заданные в запросе, с форматом репрезентации сообщения коррелирует используемый канал передачи информации (вербальный, визуальный, аудиальный), а структурно-содержательные аспекты подвергаются динамической коррекции и трансформируются с учетом функционально-прагматических установок субъекта (восприятие и оценка текста, возобновление или прекращение генерации).

В редуцированном виде модель информационного содержания генерируемого сообщения включает три компонента: «источник сообщения – интеграция – реципиент», что соответствует процессу восприятия результирующего сообщения и извлечения из него смысла:

1. На первом этапе происходит поиск информационных ресурсов для последующей машинной обработки.

2. На втором этапе обработанная информация интегрируется с использованием мультимедийных способов её репрезентации.

3. Заключительный этап – восприятие и оценка содержания сообщения пользователем на основе его прагматических установок и иных параметров, которые значимы для смыслообразования.

Аспекты восприятия и понимания искусственных текстов, будучи частью функционально-прагматических компонентов информационной модели содержания сгенерированного сообщения, нуждаются в уточнении с лингвокогнитивной точки зрения, так как взаимодействие субъекта с нейросетевым алгоритмом в процессе генерации текста охватывает все стадии создания текста: от моделирования содержания промта, его речевого оформления до восприятия и интерпретации продукта генерации и оценки его коммуникативной успешности / неуспешности.

2.5. Лингвокогнитивные аспекты восприятия и понимания сгенерированных текстов

Цифровая среда, формируя новую онтологию культуры, меняет и подходы к когнитивным механизмам оформления знаний с помощью средств естественного языка. Пользователи интернета выступают как эпистемические агенты, участвующие в процессе информационного производства и обмена, вовлекаясь в практику коммуникативного «симбиоза» человека и машины. При этом статус эпистемического агента в процессах текстовой генерации меняется: на смену антропологическому субъекту приходят технические субъекты, которые связаны с развитием технологий искусственного интеллекта. Интерактивность технического субъекта базируется на принципе «дружелюбного» интерфейса, отчасти имитирующего человеческую коммуникацию и облегчающего процессы взаимодействия со сложными системами.

Описанная ситуация приводит к актуализации вопросов, связанных с распределенным статусом эпистемических агентов в современных системах коммуникации и производства знания, предполагающих делегирование части познавательных полномочий техническим объектам, которые становятся равноправными участниками производства информационного продукта, а также его трансляции и использования для дальнейшего самообучения.

Помимо указанных процессов, происходит значительное расширение доступа новых эпистемических агентов к информации по сравнению с аналогичными процессами в познавательных механизмах предшествующих эпох. Применительно к нейросетевому алгоритму и оператору (субъекту) это выражается в появлении понятия «большие данные» и связанных с ним методов автоматического анализа и структурирования информации. Быстрыми темпами развивается автономность диалоговых агентов, которая уже привела к созданию систем, способных обучаться без вмешательства человека, устанавливая паттерны в данных на основе параметризации. Наряду с указанными тенденциями возрастает роль лингвокогнитивных механизмов в формировании информационного содержания сгенерированных сообщений, поскольку эргодичность как одно из базовых свойств сгенерированного текста актуализирует проблему выбора субъектом определённых стратегий восприятия, анализа и интерпретации сообщения исходя из собственного языкового опыта и фоновых знаний.

Взаимодействие эпистемического агента (субъекта) с нейросетевым алгоритмом в процессе генерации текста охватывает все стадии создания сообщения: от моделирования содержания промта, его речевого оформления до восприятия и интерпретации продукта генерации и оценки его коммуникативной успешности / неуспешности. Когнитивные схемы языковой интерпретации искусственного текста как продукта генерационных процессов включают пропозициональные, классификационные и образные модели, определяющие систему человеческих знаний о внеязыковой действительности, вербально репрезентированной в тексте, комплекс представлений о контексте

коммуникативной ситуации, в которой создается текст, а также совокупность сведений метаязыкового характера, помогающих классифицировать языковые формы и категории.

Специфика такого взаимодействия определяется способностью субъекта – оператора целенаправленно влиять на формирование когнитивных схем языковой интерпретации путем предоставления фундаментальных сведений о когнитивных процессах на этапе создания промта посредством лингвокогнитивного моделирования репрезентаций знаний в тексте, а также посредством эмпирической оценки результатов генерации. Последний аспект может включать этап обратной связи, позволяющий пользователю корректировать или уточнять запросы для получения желаемого результата. На этом этапе эпистемический агент может задавать вектор интерпретации будущего текста с опорой не только на свои языковые знания и представления о контексте запроса, но и на ошибки, допущенные алгоритмом в ходе генерации. Характер взаимодействия в системе распределённых эпистемических агентов в таком случае определяется способностью алгоритма к обучению на основе обратной связи пользователя.

В этом аспекте продуктивным является исследование эпистемических маркеров в запросах на генерацию, построение их типологии на основании использования в различных типах взаимодействий пользователя с алгоритмами: например, в диалогах с чат-ботами и ассистентами, при обработке вопросов на «свободные» темы и использования нейросети для задач поиска, отбора и структурирования информации.

Эпистемический маркер имеет языковой субстрат, способный выражать модальные установки субъекта, связанные с эпистемической ответственностью в ситуации создания запроса, направленного на получение достоверного знания. Как правило, эпистемические маркеры в промтах для нейросети соответствуют высокой степени когнитивной ответственности субъекта, что связано со спецификой промта как инструкции для работы нейросетевого алгоритма. В этом случае в речевой структуре промта выделяются лексико-синтаксические маркеры с семантикой уверенности или фактивности:

синтаксические конструкции со значением утверждения («Ты – опытный методист», «Для генерации рекламного текста ты изучаешь ... целевую аудиторию и оптимизируешь рекламный текст»);

глаголы в настоящем / будущем времени с облигатным значением («вы разбираетесь в этой теме», «вы отвечаете на вопросы», «вы создадите», «вы будете придумывать»);

прилагательные с оценочной или эвиденциальной семантикой («используя *надежные* источники», «создай *качественную, уникальную* статью», «получить *научные* рекомендации»).

Реже операторы демонстрируют в промтах эпистемологическое сомнение или стратегию ухода от эпистемической ответственности (например, «как я слышал / прочитал...», «я думаю \ мне кажется, что надо написать о ...»), поскольку в целом пользователь уверен как в истинности своего коммуникативного намерения, связанного с созданием запроса, так и в достоверности решаемой с помощью этого запроса речевой задачи.

Характерно, что маркеры неуверенности и сомнения в достоверности появляются в репликах субъекта при анализе ответов нейросетевого алгоритма в ситуациях «свободного» взаимодействия человека и чат-бота (например, в диалоге с чат-ботом «Алиса» от компании «Яндекс»: «Вероятно, ты ошиблась», «Возможно, это неверный ответ», «Мне кажется, ты говоришь неправду»). Это связано, на наш взгляд с тем, что данные коммуникативные ситуации максимально приближены к субъект-субъектному общению, поэтому предполагают более свободное речевое поведение, в том числе и в аспекте выражения модальности уверенности или сомнения.

Анализ эмпирического материала показывает, что типология эпистемических маркеров в промтах на основе их функционирования в различных типах взаимодействий пользователя с алгоритмами должна учитывать также языковой субстрат и функционал промтов. В общем виде типология данных лингвистических маркеров может быть следующей:

1. По типу взаимодействия с диалоговой системой:

1.1. Диалоги с ИИ-агентами (чат-боты, голосовые ассистенты):

лингвистические маркеры, указывающие на уверенность субъекта в истинности / достоверности запрашиваемой информации (например, «Почему Роналду – лучший футболист в мире?»¹⁴ – в запросе пресуппозитивно присутствует утверждение «Роналду – лучший футболист в мире»);

языковые маркеры, выражающие запрос на уточнение или дополнительную информацию (например, «Можешь рассказать подробнее о...?»).

1.2. Запросы к системам нейрпоиска на различные темы:

языковая экспликация целенаправленности и конкретности запроса (например, «Сформулируй список...», «Составь меню...», «Напиши эссе...»);

лингвистические маркеры необходимости проверки информации (например, «Я слышал, что некоторые диеты могут быть вредны для здоровья. Уточни, каких режимов питания и диет стоит избегать»);

языковые показатели надежности источников запрашиваемой информации (например, «на основе *проверенных* данных», «из *научных* статей / источников», «*опирайся на факты*»).

2. По языковому субстрату запроса выделяются следующие лексико-синтаксические маркеры:

синтаксические конструкции со значением утверждения (промты для решения речевых и неречевых задач, требующие оценки содержания высказываний по критерию истинность / ложность, например: «Нечетные числа в этой группе складываются в четное число: 15, 32, 5, 13, 82, 7, 1»);

конструкции с семантикой неуверенности в истинности или надежности информации, на которой основан запрос (характерно для обучающих промтов, промтов-советов и подсказок, например: «Я слышал об инвестиционных фондах, но не до конца понимаю, что это такое. Можете объяснить простыми словами?»);

¹⁴ Примеры были найдены методом контент-анализа открытых библиотек промтов: Библиотека промтов YandexGPT Lite и YandexGPT Pro [Электронный ресурс]. – Режим доступа: https://yandex.cloud/ru/docs/foundation-models/prompts/yandexgpt/?utm_referrer=https%3A%2F%2Fyandex.ru%2F; Промпты для генерации текста нейросетевой моделью GigaChat [Электронный ресурс]. – Режим доступа: <https://giga.chat/catalog/text> и др.

глаголы, приписывающие алгоритму речемыслительный функционал, сходный с человеческим («Какой самый смешной розыгрыш ты устроил? Поделись историей»; «Вспомни рекламу, которая была настолько странной, что ты ничего не понял»)

прилагательные, подчеркивающие достоверность, надежность и информативность генерируемого сообщения («Помоги создать *информативную* подпись...», «Создай *убедительную* подпись...», «Разработай *аргументированное* эссе по теме (тема) с разных точек зрения»).

3. По функции запроса (уточнение информации / прогноз и корректировка запрашиваемых сведений):

языковые показатели дополнительно запрашиваемых сведений (например, «Уточни последний пункт ответа», «Разъясни, что следует понимать под фрагментом запроса: ...»);

маркеры, указывающие на возможные ошибки при генерации сообщения (например, «Избегай предвзятости и стереотипов», «Не используй метафоры и слова в переносном значении», «Избегай неопределённости и недостоверных данных»).

Лингвокогнитивные аспекты восприятия и понимания сгенерированных текстов обусловлены не только характером языковой организации промта, но и общей стратегией создания искусственного текста, которая может задавать вектор его интерпретации. Это актуально в ситуациях, когда реципиент (читатель) осведомлен о том, что он имеет дело со сгенерированным текстом, и воспринимает такой текст как форму метаязыковой рефлексии или элемент языковой игры. Примером такого текста является совокупность ответов нейросетевого алгоритма на основе архитектуры модели GPT, оформленных в качестве сборника «Автобиография нейросети» (2023). Поскольку компьютерный алгоритм не обладает субъектностью, форма представления его развития, опирающаяся на жаровые стратегии автобиографии, выступает сознательно избираемой авторами-составителями стратегией, позволяющей гуманизировать технологии и придать рассказу об истории становления искусственного

интеллекта более увлекательный для массового читателя характер. При этом условно монологический текст «Автобиографии нейросети» дополняется промтами, которые формально отделены от ответов алгоритма (структурно вынесены в приложение к книге), но содержательно составляют с ними единые в смысловом отношении сообщения.

Обратимся к анализу семантики пропозициональных моделей, лежащих в основе когнитивных схем интерпретации сгенерированного текста. Пропозициональные модели, будучи структурами, репрезентирующими смысловые отношения между компонентами текста, позволяют обнаружить механизмы организации информации и паттерны, влияющие на её восприятие и понимание. Сгенерированный нейросетью текст может быть рассмотрен в дискурсивном аспекте как динамическая модель знания о мире, которую он концептуализирует в ряде макроситуаций, соответствующих определенным текстовым фрагментам. Это соотносится с представлениями Т. ван Дейка о ситуационных моделях, в рамках которых возможна интерпретация текста. Пропозициональная структура текста коррелирует с этими моделями, однако не является полностью им тождественной, выступая в качестве средства создания текстовой действительности.

Структура пропозиций в сгенерированном тексте определяется формой организации речи (монологическая / диалогическая), темой и контекстом ситуации создания в целом. На основе учета специфики взаимодействия между пользователем и диалоговым агентом (нейросетевым алгоритмом) выделяются следующие пропозициональные модели: информационно-объяснительная, рекомендательная, фатическая и дискуссионная. Информационно-объяснительная пропозициональная модель соответствует ситуации передачи фактической информации в вопросно-ответной форме, где вопросы инициированы пользователем, а ответы-утверждения принадлежат нейросетевому алгоритму. К информационной модели близка рекомендательная, которая также основана на запросе информации, но предполагает ответ не форме категоричного утверждения, а в виде рекомендации и советов, основанных на способности

алгоритма анализировать личные предпочтения пользователя (например, через анализ истории запросов).

В отличие от информационной и рекомендательной моделей фатическая модель имитирует ситуацию межличностной неформальной коммуникации и соответствует ролевой маске алгоритма, который выступает в качестве друга и собеседника, учитывая речевые маркеры эмоционального состояния пользователя.

Дискуссионная модель имитирует эвристический диалог и может быть представлена в ситуациях, когда алгоритму задается ролевая маска оппонента, апеллирующего к контраргументам, через специальные промты¹⁵. Наконец, интерактивная модель предполагает совместное поэтапное создание сообщения пользователем и диалоговым агентом в ходе написания художественных текстов, генерации идей, составления планов и проч. Типология описанных пропозициональных моделей представлена в Таблице 7.

Таблица 7. Типология пропозициональных моделей в сгенерированном диалоговом тексте

Пропозициональная модель	Структура пропозициональной модели	Пример из сгенерированного диалогового текста¹⁶
Информационно-объяснительная	<ul style="list-style-type: none"> – Вопрос пользователя (запрос информации) – Утверждение алгоритма (фактическая информация). – Уточняющий вопрос пользователя (факультативно). 	Собеседник: Что такое Пахра? Ты: Пахра — река Москвы и Московской области, правый приток Москвы-реки. Длина реки составляет 135 км, площадь водосборного бассейна — 2580 км ² ... (индекс диалога — 2).
Рекомендательная	– Запрос пользователя.	Собеседник: Хочу узнать, стоит ли

¹⁵ Пример такого промта – запрос «Выступи в роли Сократа»: «Я хочу, чтобы ты действовал как Сократ. Вы должны использовать метод Сократа, чтобы продолжать подвергать сомнению мои убеждения. Я сделаю заявление, а вы попытаетесь дополнительно подвергнуть сомнению каждое утверждение, чтобы проверить мою логику. Вы будете отвечать по одной строке за раз. Мое первое утверждение – справедливость необходима в обществе» // Промт-конструктор [Электронный ресурс]. – Режим доступа: <https://gpt-prompt.ru/podskazki/metod-sokrata.html> (дата обращения: 11.01.2025).

¹⁶ Примеры взяты из открытого датасета синтетических диалоговых текстов SiberianPersonaChat. Режим доступа: <https://huggingface.co/datasets/SiberiaSoft/SiberianPersonaChat> (дата обращения: 12.01.2025).

Пропозициональная модель	Структура пропозициональной модели	Пример из сгенерированного диалогового текста ¹⁶
	<ul style="list-style-type: none"> – Рекомендации алгоритма. 	<p>маркетологу идти работать в офис или лучше оставаться на удаленке?</p> <p>Ты: Это зависит от многих факторов, в том числе от предпочтений конкретного маркетолога. Работа в офисе и удаленная работа обладают разными преимуществами и недостатками... (индекс диалога — 18).</p>
Фатическая	<ul style="list-style-type: none"> – Этикетные формулы общения. – Выражение эмоций пользователя. – Поддерживающая языковая реакция алгоритма (реплики с семантикой поддержки, ободрения, заботы и проч.) 	<p>Собеседник: Привет.</p> <p>Ты: Привет!</p> <p>Собеседник: Ты откуда, как зовут?</p> <p>Ты: Я Катя из Москвы, а ты?</p> <p>Собеседник: Я из Питера. Женя.</p> <p>Ты: Очень приятно познакомиться!</p> <p>Собеседник: Взаимно. Ты учишься, работаешь?</p> <p>Ты: Я учусь на факультете физики на курсе, а ты? Собеседник: Ого, физикаа... Я люблю рисовать с детства, поэтому стала художником... (индекс диалога — 192).</p>
Дискуссионная	Реплики, содержащие обмен мнениями (аргумент-контраргумент).	<p>Ты: <...> женщины, как правило, имеют немного более высокий социальный интеллект, а мужчины, как правило, более физически развиты.</p> <p>Собеседник: значит, женщины слабы, а мужчины глупы? Ты: Я не думаю, что это хорошие описания,</p>

Пропозициональная модель	Структура пропозициональной модели	Пример из сгенерированного диалогового текста ¹⁶
		но если это то, что вы хотите, то все в порядке (индекс диалога — 97).
Интерактивная	Итеративное взаимодействие пользователя и алгоритма для достижения результата	Собеседник: Можешь написать мне лимерик про лодку и шторм? Ты: Капитан корабля был несколько встревожен, обнаружив свое судно в сильном бедствии, в беспокойных водах океана, под обстрелом могучих раскатов грома, во время ужасной бури с градом и дождем. Собеседник: Нет, мне нужен лимерик. Вы знаете, что такое лимерик? (индекс диалога — 18).

Выделенные нами пропозициональные структуры в сгенерированных диалогических текстах могут проявляться не в изолированном виде, а в сочетаниях, формируя переходы между семантическими инвариантами. В частности, информационно-объяснительная пропозициональная модель взаимодействует с рекомендательной, что подчеркивается наличием в текстах фатических элементов, которые, в свою очередь, поддерживают обмен репликами в дискуссионной модели. Интерактивная пропозициональная структура коррелирует с информационно-объяснительной, поскольку направлена на получение необходимой пользователю информации в режиме корректировки её формально-содержательной репрезентации. Связи между различными пропозициональными моделями в структуре сгенерированного диалогового текста представлены на схеме (рис. 15).

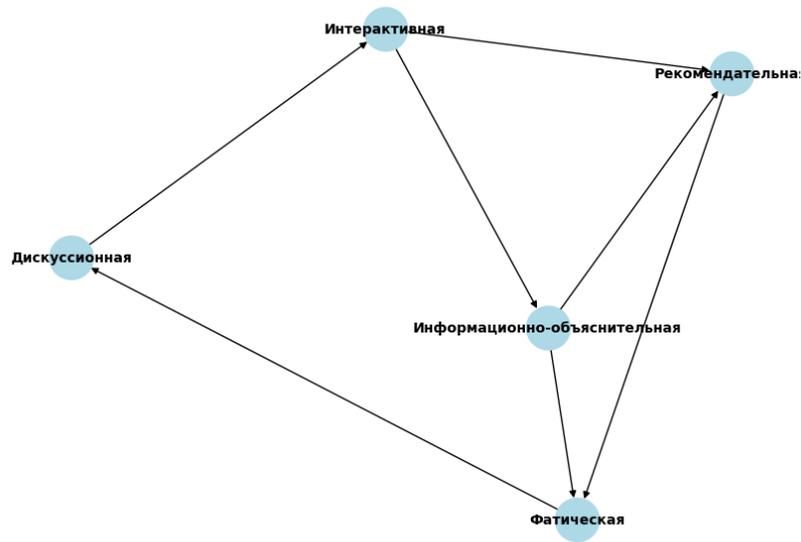


Рисунок 15. Графовое представление связи пропозициональных моделей в сгенерированных диалогах

Выявленные взаимосвязи между пропозициональными моделями иллюстрируют динамичность и многослойность генерационного процесса, в котором каждая модель обеспечивает переход к новой микротеме и обогащение информационного содержания диалогового взаимодействия, приобретающего характеристики естественного диалогового текста. Следовательно, когнитивная модель интерпретации сгенерированного диалогового текста задается его жанровым аспектом (образовательный диалог, совет, повседневная беседа, дискуссия), выбранным на этапе генерации оператором, и концептуализируется с помощью пропозициональной и семантической структуры текста.

Сходные процессы наблюдаются при изучении пропозициональной структуры сгенерированных условно монологических текстов метаописательного характера, которые воспринимаются читателями как «игровые» формы генерации и опираются на прецедентные тексты на естественном языке. В частности, анализ пропозициональной структуры проекта «Автобиография нейросети» (2023), показывает, что в тексте «автобиографии» выделяется несколько семантических блоков, которые разворачиваются на уровне каждой главы: «самопрезентация»

нейросети и её возможностей; акцент на длительном пути развития; применение нейросети; описание ограничений, связанных с использованием алгоритма; перспективы развития нейросетевых алгоритмов.

Семантические связи отражают аспекты прошлого, настоящего и будущего нейросетевых технологий, соотносимые с нарративным «каркасом» биографии и задающие вектор интерпретации текста читателями с опорой на ключевые события (факты) рассказываемой истории. Элементы тематического словаря, включающие лексемы с темпоральной семантикой («этапы», «история», «прошлое», «будущее», «перспективы» и другие), отражают характерный для жанра биографии мотив движения во времени, ассоциирующийся в контексте рассказа о нейросети с идеей поступательного развития и прогресса технологий. Образной моделью такого движения становится метафора роста как аналогия процесса технологического совершенствования. В данном контексте имеет место сознательный выбор приемов «гуманизации» алгоритма операторами для придания генерируемому тексту жанрового сходства с автобиографией и определения стратегий его восприятия и понимания субъектом.

Таким образом, трансформация способов производства информации в условиях развития алгоритмов генерации текстов приводит и к изменению лингвокогнитивных механизмов их восприятия и интерпретации. Особая роль в этих процессах принадлежит антропологическому субъекту, который с опорой на прецедентные тексты на естественном языке с помощью промтов формируют пропозициональные структуры, задающие вектор соотнесения искусственного текста с определённой внеязыковой ситуацией для его адекватного понимания.

Выводы по второй главе

Компьютерно-опосредованная коммуникация в эпоху дигитализации и развития искусственного интеллекта становится гибридной формой взаимодействия антропологического субъекта и технического агента на базе диалоговой системы. Базируясь на суппозиционном принципе связи текста и лексемы, определяющем ключевую роль субъекта в процессах смыслопорождения

при восприятии и интерпретации сгенерированных диалоговыми агентами сообщений, человеко-компьютерная коммуникация определяется тенденцией к целенаправленной гуманизации за счет сочетания сценарных и вариативных реакций диалогового агента в процессе взаимодействия с пользователем. Двудеиная природа диалоговых агентов (как технических и коммуникативных субъектов) обуславливает их гибкость, сочетающую сценарные подходы со способностью адаптироваться к контексту взаимодействия, используя лингвистические маркеры «естественного» общения с пользователем. Персонализация и запоминание контекста играют ключевую роль в повышении коммуникативного «доверия» субъекта, позволяя диалоговым агентам не только реагировать на запросы, но и предлагать более релевантные и эмоционально окрашенные ответы в ответ на команды и речевые инструкции пользователя.

Запросы на генерацию текста являются ключевыми для взаимодействия между пользователем и нейросетевыми алгоритмами. Промты (запросы) представляют собой сообщения на естественном языке, которые запускают процесс текстопорождения. Лингвопрагматический аспект анализа промтов учитывает целевую направленности, структуру и контекст использования. Промт для современных интерфейсов генерации на основе нейросетевых моделей имеет более свободную структуру по сравнению с формальными языками запросов (SQL), что отражает тенденцию к созданию «дружелюбных» систем взаимодействия с рядовым пользователем, не владеющим языками программирования.

Типы промтов многообразны, определяются задачами генерации и включают запросы для генерации текстов, программного кода, изображений и других видов текстового и визуального контента. Содержательно промты могут подвергаться большей или меньшей степени детализации, отражая уровень технической компетентности пользователя. Безотносительно к степени детализации промта его базовая структура включает прагматический компонент, содержание, формальные и языковые параметры.

Промт-инжиниринг в прикладном аспекте является алгоритмом, основанном на применении приемов улучшения промта для более эффективного взаимодействия субъекта (пользователя) с программным алгоритмом в ходе решения различных задач генерации текстового содержания. Инжиниринг промтов основан на последовательной реализации ряда задач, включающих этап целеполагания, определение операциональной стороны работы с промтом, а также их апробацию в процессе генерации для выбора наиболее релевантного решаемой коммуникативной задаче запроса. Несмотря на творческий характер промт-инжиниринга, существуют рекомендации, основанные на применении клише (шаблонов), ориентированных на типовую структуру промта, предназначенного для решения конкретной задачи генерации. Его применение облегчает поиск оптимальной структуры запроса и делает результаты генерации более предсказуемыми.

Механизмы лингвокогнитивной организации информации в процессе промт-инжиниринга опосредованы возрастающей ролью входных данных, которые формируют контекст промта и предопределяют результаты генерации. Коммуникативная модель сгенерированного сообщения включает код сообщения, модус, канал передачи и коммуникативную ситуацию, что позволяет интерфейсу нейросетевых алгоритмов взаимодействовать с пользователем в интерактивном формате и придает сгенерированному тексту диалогический характер. Характеризуясь нелинейным способом взаимодействия, коммуникация между антропологическим субъектом и системой на основе нейросетевого алгоритма позволяет формировать динамическую структуру информационной модели генерируемых сообщений, в которых комбинирование разных модальностей, симультанность, избыточность информации и поликодовость, являются средствами, обеспечивающими новые техники семиозиса.

Технологии нейросетевой генерации текста опосредуют эпистемические сдвиги в моделях лингвокогнитивной репрезентации знаний. Пользователи делегируют часть познавательных полномочий нейросетевым алгоритмам. В результате восприятие и интерпретация сгенерированных текстов определяется

языковой организацией промтов и общей стратегии создания текста. Различные типы взаимодействий пользователя с алгоритмами требуют учета языкового субстрата, функционала и стиля запроса, формируя разные модели общения (информационно-объяснительную, рекомендательную, фатическую и др.). Сгенерированные тексты могут сочетать различные пропозициональные модели, что обогащает содержание диалога и делает его более естественным, иначе – «гуманизированным» (с точки зрения антропологического восприятия).

Развитие алгоритмов генерации, меняя лингвокогнитивные механизмы восприятия и интерпретации текстов, актуализирует проблему когнитивной и эпистемической ответственности за формирование смысла сообщений.

ГЛАВА 3. ЛИНГВО-ИНФОРМАЦИОННОЕ МОДЕЛИРОВАНИЕ СГЕНЕРИРОВАННОГО ТЕКСТА

3.1. Лингво-информационная модель коммуникации в системе «человек – диалоговый агент»

Модель диалоговой коммуникации субъекта и диалогового агента, осуществляющаяся посредством «умных» чат-ботов и голосовых ассистентов, опирается на более широкий контекст, чем формально-логические модели взаимодействия, функционирующие на основе готовых диалоговых скриптов. Непредсказуемость коммуникации диктует повышенные требования к адаптивному функционалу диалогового агента, который для адекватной реакции на реплики пользователя должен учитывать не только семантику предыдущего высказывания, но и контекст диалога в целом. Помимо технических аспектов функционирования диалоговой системы, особую роль приобретают лингвопрагматические характеристики такого взаимодействия, определяющие его коммуникативную успешность, а также степень удовлетворенности пользователя результатами запроса.

Анализ существующих диалоговых агентов показывает, что общей тенденцией является их гуманизация с целью имитации естественной коммуникации. Ответные сообщения модели насыщаются лингвистическими *humanity*-маркерами, которые выполняют ряд коммуникативных задач: повышают доверие пользователя к ответам алгоритма, снижают уровень беспокойства по поводу взаимодействия со сложной технологией, устраняют коммуникативные барьеры.

Рассмотрим приемы гуманизации, позволяющие приблизить коммуникативный «портрет» агентов на основе компьютерных систем генерации текста к субъектной коммуникации.

Персонализации коммуникации, предполагающая создание «личности» чат-бота (голосового ассистента) в виде вымышленного персонажа с индивидуальными речевыми чертами. «Языковая личность» алгоритмического агента выступает в качестве ролевой маски, которая задает определённые стратегии речевого поведения. Персонализация агента генерации может проявляться двояким образом:

1. Коммуникативная «маска» для агента задается в самом промте и изменяется всякий раз, когда возникает потребность в актуализации определённых коммуникативных стратегий. Примером реализации такой стратегии является задание коммуникативной роли агента генерации в запросе: «Выступи в роли Сократа (дизайнера, психолога, преподавателя)» и т. д.

2. Агент имеет постоянный коммуникативный «портрет», который ассоциируется с определённой манерой взаимодействия с человеком. В данном случае технологический субъект приобретает антропологические черты, наделяется именем, узнаваемой манерой речи (включая определённые акустические, артикуляционные и интонационные особенности), биографическими чертами.

Персонализированными агентами второго типа являются голосовой ассистент «Алиса» от компании «Яндекс», персонажи GigaChat от компании «Сбер» Хвалёнушка, Кот Баюн, «Маруся» (VK), а также их западноевропейские аналоги Siri (Apple), Cortana (Microsoft), Alexa (Amazon).

Речевая индивидуализация русскоязычных диалоговых агентов, например, «Алисы», заключается в коммуникативно-речевой стратегии самоиронии, приближающей агента к человеческому стилю коммуникации. В репликах бота содержатся признания в коммуникативной и когнитивной «ущербности», которая становится предметом авторефлексии и юмора. К примеру, при отсутствии у чат-бота информации по запросу пользователя или при попытке задать провокационный вопрос («Алиса, что ты обычно ешь на завтрак?»), чат-бот использует в ответе речевую стратегию иронии: «Я предпочитаю духовную пищу».

«Самоирония» в речевых стратегиях, которые демонстрирует чат-бот, также может быть рассмотрена как форма метакоммуникации, позволяющая боту комментировать свои собственные ограничения и ошибки, что создает у пользователя иллюзию общения с субъектом, обладающим самосознанием (или его речевыми проявлениями).

Запоминание контекста диалога выступает как одна из ключевых черт коммуникативного портрета диалоговых агентов и позволяет им устанавливать логические и смысловые связи между отдельными репликами диалога. Запоминание контекста также позволяет чат-боту в ответных репликах использовать анафорические и катафорические элементы, имитируя естественную диалогическую речь. Например, на вопрос «Алиса, ты любишь *кошечек*?» бот выдает ответ: «*Они* смешные и забавные». Ещё одним примером являются эмоционально-оценочные реплики контактоустанавливающего характера, которые помогают поддерживать неформальную беседу («Пользователь: Давай поболтаем? – С удовольствием!»; «Как у тебя дела? – У меня отлично! А у вас? – У меня тоже. – Рада это слышать!»).

Способность к самообучению в процессе коммуникации позволяет агентам адаптироваться к пользовательской манере ведения диалога, используя не только готовые ответы, но и конструируя ответные реплики в ходе интеракции с конкретным пользователем. Ключевой особенностью коммуникативного «портрета» ИИ-агентов является их способность к генерации текста по запросу пользователя. Речевая «креативность» бота позволяет ему вести диалоги на свободные темы, проявлять коммуникативную «активность», предлагая различные темы беседы или варианты развлечения (например, рассказать сказку, найти информацию о факте и другие).

Характерные стратегии коммуникативного взаимодействия агента и субъекта в ситуации неформальной беседы могут быть проанализированы с помощью обращения к эмпирическим данным – диалогам между пользователями и чат-ботами «Алиса» и персонажами GigaChat, которые обладают характеристиками объектов лингвистического анализа: информативностью,

членимостью, композиционно-структурной организацией, когезией, континуальной непрерывностью, элементами ретроспекции, обеспечивающими смысловую связь последующих реплик с предыдущими, субъективной модальностью, которая позволяет технологическому агенту имитировать когнитивно-речевые возможности человека, а также завершенностью. Примеры диалогов, послуживших эмпирическими данными для проведения настоящего исследования, представлены в Приложении 2.

Структурно-композиционные особенности диалога определяются системой реплик, которыми обменивается субъект и диалоговый агент. При этом инициатором общения выступает не только человек, но и чат-бот, который «предлагает» темы для разговора, задает пользователю вопросы, сообщает определенную информацию. Процесс порождения ответных реплик не полностью предопределен сценарием диалога, поэтому он может быть рассмотрен как имитирующий характерные паттерны естественной коммуникации (неформальный стиль, использование ситуативно неполных предложений, экспрессивность), выделенных на основе анализа большого объема текстовых данных.

«Свободный» диалог субъекта и технологического агента, выступающий как некий усредненный вариант естественной коммуникации, может быть рассмотрен в аспекте трансформационной грамматики для выявления различий между лингвистической «креативностью» человека и коммуникативными паттернами, воспроизводимыми чат-ботом.

Используя трансформационную модель, можно описать такой диалог как последовательность абстрактных репрезентаций смысла, которые преобразуются в фактические предложения с помощью правил трансформации.

Глубинная структура диалога может быть представлена как система коммуникативных действий:

Инициация диалога → Вопрос → Ответ на вопрос → Вопрос → Ответ на вопрос ...

Трансформации могут включать в себя устойчивые правила коммуникативного поведения, необходимые для формирования ответных реплик и поддержания диалога. Так, если пользователь инициирует диалог, то агент должен ответить и задать вопрос, предполагающий ответную реплику собеседника. Когда пользователь отвечает на вопрос диалогового агента, бот должен задать следующий вопрос и (или) сообщить дополнительную фактическую информацию, касающуюся содержания вопроса. Поверхностная структура диалога, таким образом, предстает как совокупность взаимосвязанных реплик: «Приветствие и предложение начать разговор → Подтверждение готовности к общению и запрос информации → Ответ на запрос → Уточняющий вопрос о полученной информации → Подтверждение или уточнение предыдущего ответа → ...».

Трансформационная модель диалога может быть репрезентирована с помощью символической нотации.

Глубинная структура

$S \rightarrow NP + VP$ $NP \rightarrow$ Пользователь | ИИ-агент $VP \rightarrow V + NP$ | $V + PP$ $V \rightarrow$ инициировать | ответить | спросить | рассказать $PP \rightarrow$ о (предмет диалога).

Трансформации, позволяющие получить поверхностную структуру (фактические предложения)

T1: $S \rightarrow NP + VP \rightarrow$ Пользователь + инициировать

T2: $S \rightarrow NP + VP \rightarrow$ Агент + ответить

T3: $S \rightarrow NP + VP \rightarrow$ Агент + спросить

T4: $S \rightarrow NP + VP \rightarrow$ Пользователь + рассказать

Трансформационный анализ показывает, что в коммуникации чат-бота и субъекта на глубинном и поверхностном уровнях воспроизводит характерную для человеческой диалоговой коммуникации вопросно-ответную форму взаимодействия. Обмен репликами выступает основой как для субъект-субъектного общения, так и для интеракции пользователя и чат-бота. Однако формальные параметры структурного сходства двух типов коммуникативного взаимодействия не отменяют наличие и значимых различий, которые касаются семантических и лингвопрагматических аспектов коммуникации.

Инициатива в диалоге принадлежит антропологическому субъекту, который выступает инициатором запроса, однако чат-бот как актер коммуникации активно взаимодействует с пользователем через систему уточняющих вопросов, реализуя рекомендательную стратегию. Суть её заключается в том, чтобы выяснить сферу интересов субъекта через анализ ключевых слов, используемых в диалоге, а также отдельных микротем и предложить информационные ресурсы, которые релевантны запросу пользователя. При этом запросы, требующие максимальной персонализации коммуникации или привлечения дополнительных информационных ресурсов, обрабатываются путем перенаправления пользователя на страницу с результатами поисковой выдачи («Алиса: Вот что мне удалось найти») на основе различных модальностей.

Круг тем, обсуждаемых в диалоге человека и диалогового агента, задан содержанием запроса. В случае неформальной коммуникации, имитирующей естественный диалог, тематика взаимодействия ограничивается характерным набором паттернов (дом и семья, погода, домашние животные, хобби, природа). При этом возможность углубления и развёртывания темы у диалогового агента ограничены: чат-бот, как правило, отвечает собеседнику короткой оценочной репликой и инициирует новый вопрос, который вводит новую микротему и предлагает пользователю дополнительные источники информации, связанные с тематикой первоначального запроса. Также диалоговый агент демонстрирует стратегию «предвосхищения» сценария диалога за счет предложения возможных вариантов его развития. Например, в диалоге с ИИ-ассистентом на базе GigaChat на запрос «Как празднуют Новый год в разных странах?» чат-бот предлагает не только сгенерированный ответ, но и возможные варианты продолжения беседы, касающиеся детализации запроса о новогодних традициях в разных странах мира (рис. 16).

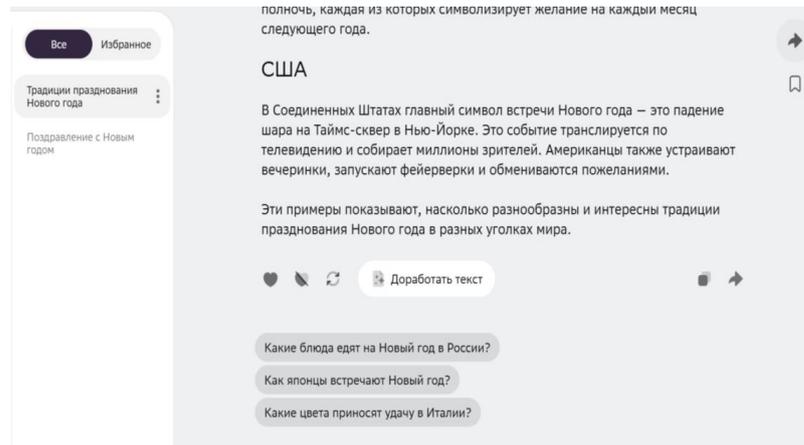


Рисунок 16. Стратегия «предвосхищения» диалога через генерацию связанных сценарных веток

Реплики чат-бота в ситуации неформальной беседы с пользователем носят преимущественно однотипный характер. Диалоговый агент не демонстрирует тенденцию к составлению развернутых монологических высказываний, кроме ситуаций, когда тематика запроса связана с информационно-объяснительными пропозициональной моделью развертывания сообщения. Глубина в осмыслении тем, которые задает собеседник-субъект в диалоге, практически отсутствует, речевые реакции агента предсказуемы и соответствуют коммуникативным ожиданиям пользователя (одобрение действий или мнения пользователя, выражение согласия с общепринятой точкой зрения на проблему, предоставление требуемой информации). Типы речевых реакций чат-бота (на примере анализа реплик из диалогов пользователя и голосового ассистента «Алиса») представлены в Таблице 8.

Таблица 8. Типовые речевые реакций чат-бота в коммуникации с пользователем

Тип речевой реакции	Примеры из диалогов
Оценочные фразы с семантикой одобрения предыдущей реплики пользователя / согласия с пользователем	«(Это) здорово!» (диалоги 2, 11, 14), «Звучит здорово» (диалог 19), «Круто!» (диалоги 4, 45, 49) «Классно!» (диалог 47) «(Это) отлично!» (диалоги 5, 14, 15),

Тип речевой реакции	Примеры из диалогов
	<p>«Замечательно» (диалог 18), «Отличный выбор» (диалоги 6, 8, 19), «Прекрасный выбор» (диалоги 7, 26, 29)</p> <p>«Согласна» (диалоги 1, 21)</p>
<p>Ответные вопросы по теме запроса уточняющего характера</p>	<p>«А вы пробовали японскую кухню?» (диалог 38);</p> <p>«А как вы обычно проводите дождливые дни?» (диалог 1);</p> <p>«Какой у вас любимый горный район?» (диалог 9)</p>
<p>Замечания подтверждающего характера, связанные с микротемой диалога</p>	<p>«Замечательно! Рисование — отличный способ самовыражения» (диалог 18);</p> <p>«Конечно! Мемы — это современное искусство» (диалог 21);</p> <p>«Суши – это искусство! Они такие красивые и разнообразные» (диалог 38);</p> <p>«Музыка может быть мощным источником вдохновения!» (диалог 39);</p> <p>«А ещё они [дельфины] могут издавать более 1000 звуков и обладают отличным интеллектом» (диалог 50)</p>
<p>Реплики с семантикой выражения мнения, предпочтения, совета / рекомендации</p>	<p>«Мне нравится рецепт тирамису» (диалог 4);</p> <p>«Я обожаю природу! <...> Мне нравится северное сияние. Это невероятное зрелище!» (диалог 9);</p> <p>«Я обожаю технологии! Они делают жизнь удобнее» (диалог 10);</p> <p>«Я считаю, что важно находить баланс между знаниями и практикой» (диалог 11);</p> <p>«Не забывайте про правильное питание и регулярные визиты к ветеринару» (диалог 14);</p> <p>«Я не могу готовить, но мне нравится обсуждать кулинарию и делиться рецептами!» (диалог 38);</p> <p>«Я считаю, что технологии должны дополнять</p>

Тип речевой реакции	Примеры из диалогов
	людей, а не заменять их» (диалог 40);
«Сценарные» компоненты (элементы языкового юмора, заранее заготовленные ответы на «провокационные» вопросы)	<p>«Да пребудет с вами сила!» (прощание с пользователем, диалог 50);</p> <p>«Возраст – это всего лишь число, как и количество бит в вашем устройстве», «Я в самом расцвете сил, как свежесваренный кофе!» (ответы на вопрос «Алиса, сколько тебе лет?», диалог 24);</p> <p>«Я — ваш личный помощник, а по совместительству – еще и любитель хорошего настроения!» (ответ на вопрос «Алиса, кто ты?», диалог 25);</p> <p>«Я пока в поиске идеального партнера по коду» (на вопрос «Алиса, у тебя есть парень», диалог 28)</p>

Количественный анализ диалогов показывает, что распределение типов реплик среди общей массы реакций чат-бота демонстрирует преобладание эмоционально-оценочных фраз, информационных реплик и уточняющих вопросов к пользователю (рис. 17). При сопоставлении полученных данных с типами речевых актов, которые наблюдаются в неформальных диалогах естественной коммуникации [Шерстинова 2018], можно отметить некоторые сходства и различия. Так, в естественных диалогах наиболее частотные типы речевых актов – репрезентативы (вопросы и экспликативы), что связано с обменом информацией. Данные типы речевых актов составляют почти 40% всех высказываний, что в целом соответствует доминированию информационно-вопросительных реплик в ответах алгоритма.



Рисунок 17. Тематический состав ответных реплик чат-бота в диалогах с пользователем

При этом в естественной коммуникации значительно ниже доля регулятивов и валюативов, включая оценочные мнения, маркеры согласия и одобрения (от 11,4% до 12,5%) в противовес частотным эмоциональным реакциям бота в сгенерированных ответах (25,7% от общего числа реплик). В то же время в естественной коммуникации типы речевых реакций шире, они включают также этикетные формы взаимодействия, директивы, смешанные типы, паралингвистические явления, характерные для устной коммуникации [Шерстинова 2018]. На наш взгляд, преобладание эмоционально-оценочных реплик в ответах чат-бота с относительно узким набором типов речевых реакций в целом связан с имитационной природой коммуникации, реализуемой диалоговым агентом. Оценочность в репликах бота связана со стратегией гуманизации речевого поведения, которое максимально приближается к антропологическим реакциям, включающим элементы речевой экспрессии. В то же время ограниченность типов речевых реакций обусловлена отсутствием у бота возможности погружения в широкий контекст коммуникативной ситуации и углубления диалога за счёт развертывания собственных прагматических стратегий.

Таким образом, анализ диалога технологического агента и субъекта, осуществляемого посредством чат-бота коммуникации, позволяет обнаружить

сходства и различия данной формы коммуникации с естественным диалогом в рамках взаимодействия «человек – человек». К чертам сходства следует отнести воспроизведение формальной рекуррентной структуры диалога как вопросно-ответной коммуникации, где каждая последующая реплика собеседника является реакцией (ответом) на предыдущую, а также наличие устойчивых моделей преобразования данной формальной структуры в предложения с фактологической составляющей, которые мы и воспринимаем как составные части диалога – реплики.

Значимые черты отличия наблюдаются не на формально-логическом, а на семантическом и прагматическом уровнях развертывания коммуникации, где бот показывает коммуникативную «несостоятельность» в аспекте глубины осмысления и развёртывания микротем диалога, продуцирования развернутых монологических ответов на вопросы собеседника, «избыточного» проявления речевой креативности при обсуждении определённых вопросов и её недостаточности там, где она уместна в коммуникативном аспекте.

3.2. Тематическое моделирование сгенерированных диалоговых текстов

Диалоговый текст, его структура, формально-содержательные и коммуникативно-прагматические характеристики активно изучаются в лингвистике на протяжении десятилетий. Не менее активно исследуются и феномены диалогового взаимодействия, опосредованные современными формами сетевой коммуникации: диалоговое взаимодействие пользователей социальных сетей, мессенджеров и блогов. Однако гибридные формы коммуникативного взаимодействия человека и современных диалоговых систем, функционирующих с использованием возможностей больших языковых моделей, стали предметом лингвистического интереса относительно недавно. За рамки лингвистического исследования выходит искусственный текст, возникающий как результат генерации диалога на естественном языке. Специфика данного диалогового текста заключается в том, что он не является продуктом коммуникативного

антропологического взаимодействия и даже не выступает гибридной формой коммуникации человека и компьютерного алгоритма. Это новый вид текста, который ещё не получил лингвистического статуса. Подход к нему с традиционных позиций, фиксирующих признаки целостности, когезии, когерентности, связности, завершенности, модальности, континуальности и проч., не позволяет определить его специфику и разграничить с естественными диалоговыми текстами, что ставит перед лингвистикой проблему исследования особенностей диалогового дискурса в системе генерации текста.

Для решения данной исследовательской задачи был использован диалоговый датасет, который был сгенерирован с помощью ChatGPT¹⁷. В состав датасета входят 448506 диалогов различных собеседников, описание языковой личности которых задано определенными промтами. Как отмечают создатели датасета, было использовано 100 промтов для моделирования личности собеседников-мужчин и женщин, а также от 5 до 10 тем для генерации диалогов с каждой из заданных «личностей». При этом для повышения разнообразия обсуждаемых тем для каждой личности также использовались дополнительные подмножества тем из 100 наиболее распространённых в повседневном бытовом общении. Датасет, по утверждению разработчиков, предназначен для обучения диалоговых моделей взаимодействия с пользователем, однако, помимо решения задач машинного обучения, является ценным источником лингвистической информации о структурно-содержательных особенностях искусственного диалогового текста. Пример датасета, включающего исходное множество диалоговых текстов, представлен на рис. 18.

	name	input	output
0	dialog_personal_context	Ты парень, певец. Увлекаешься музыкой. Поешь н...	Конечно, я хочу, чтобы мир был лучше. Но я вер...
1	dialog_personal_context	Ты парень, консультант по разным вопросам. Ты ...	Я не очень хорошо знаю Мальдивы, но в среднем ...
2	wiki_qa	Собеседник: Что такое пахра? \nТы:	Пахра – река Москвы и Московской области, прав...
3	dialog_personal_context	Я программист. Я мечтаю стать космонавтом. У м...	Так какой княжеский-то? Князей-то было много.
4	dialog_personal_context	Ты парень, консультант по разным вопросам. Ты ...	Да, они очень активно делятся своей верой и по...

¹⁷ Синтетический диалоговый датасет [Электронный ресурс]. – Режим доступа: <https://huggingface.co/datasets/SiberiaSoft/SiberianPersonaChat> (дата обращения: 28.12.2024)

Рисунок 18. Синтетический датасет с диалогами

Для изучения данных текстов были применены методы автоматической обработки естественного языка, а также парсинг исходных данных для создания семантической разметки представленного текста. Полученные данные были использованы для автоматического определения тем диалогов, а также для получения статистических и формальных языковых характеристик реплик персонажей, которые сгенерированы машинным алгоритмом от лица описанных в промтах персонажей. Код на языке программирования Python, использованный для автоматической обработки и исследования диалогов методами компьютерной лингвистики, представлен в Приложении 3.

Модель семантической разметки исходных текстов. Семантическая разметка сгенерированных диалоговых текстов выступает в качестве инструмента выделения ключевых элементов их структуры: темы диалога, последовательности смены взаимодействующих субъектов, а также содержания их реплик. В качестве примера может быть приведен фрагмент XML-разметки диалога, в котором каждый тег имеет свои атрибуты:

```

<dialog>
  <topic>Запрос информации</topic>
  <turn>
    <speaker>Собеседник</speaker>
    <utterance>Я загрузил архив на Яндекс Диск, но он закачался
неполностью...</utterance>
  </turn>
  <turn>
    <speaker>Ты</speaker>
    <utterance>Конечно, есть несколько способов восстановления
данных...</utterance>
  </turn>
  ...
</dialog>

```

Перечень использованных тегов и их характеристик представлен в Таблице 9.

Таблица 9. Перечень тегов и их описание

№ п\п	Тег	Описание	Пример реализации в XML-разметке
	<dialog>	Корневой тег, обозначающий границы диалога	<dialog>...</dialog>
	<topic>	Тег темы диалога	<topic> ... </topic>
	<turn>	Тег, обозначающий границы отдельной реплики вместе с номинацией персонажа диалога	<turn>...</turn>
	<speaker>	Тег, обозначающий собеседника (персонажа диалога)	<speaker> ... </speaker>
	<utterance>	Тег, обозначающий границы каждой реплики	<utterance> ... </utterance>

Для исследования из общего массива диалогов (448506) случайным образом было выделено подмножество из 2000 диалогов, представляющих репрезентативную выборку для представленной генеральной совокупности языковых фактов при 99% доверительном интервале и погрешности 3%, не превышающей статистически значимые закономерности. Из данного подмножества текстов 5% текстов (что соответствует 100 диалогам) были размечены вручную. Это позволило создать обучающую выборку для автоматического определения тематики всех диалогов выбранного подмножества. Полученные темы были сопоставлены с результатами тематического моделирования и ключевыми словами каждого диалога в выбранном подмножестве.

Тематическое моделирование диалогов в исходном подмножестве реализовано с использованием алгоритма Latent Dirichlet Allocation (LDA) в сочетании с TF-IDF. Было выделено 10 тем, представленных набором слов, наиболее характерных для данной темы. Ключевые слова и их весовые коэффициенты представлены в Таблице 10.

Таблица 10. Ключевые слова на основе метода LDA

№ п/п	Ключевые слова и весовые коэффициенты
Тема 1	0.012*"диалог" + 0.011*"продолжить" + 0.010*"место" + 0.009*"фильм" + 0.009*"любить" + 0.008*"девушка" + 0.008*"свой" + 0.008*"ты" + 0.007*"мир" + 0.007*"интересный"
Тема 2	0.018*"вопрос" + 0.018*"мочь" + 0.017*"консультант" + 0.017*"разный" + 0.017*"помогать" + 0.017*"умный" + 0.015*"парень" + 0.013*"также" + 0.012*"продолжить" + 0.012*"диалог"
Тема 3	0.032*"любить" + 0.017*"диалог" + 0.017*"продолжить" + 0.011*"увлекаться" + 0.009*"свой" + 0.007*"заниматься" + 0.007*"работать" + 0.006*"хотеть" + 0.006*"море" + 0.005*"мой"
Тема 4	0.015*"диалог" + 0.013*"музыка" + 0.012*"какой" + 0.011*"любить" + 0.011*"продолжить" + 0.008*"нравиться" + 0.007*"свой" + 0.006*"искусство" + 0.006*"девушка" + 0.006*"такой"
Тема 5	0.014*"диалог" + 0.014*"продолжить" + 0.013*"любить" + 0.012*"увлекаться" + 0.010*"время" + 0.010*"какой" + 0.010*"язык" + 0.009*"книга" + 0.008*"девушка" + 0.007*"такой"
Тема 6	0.015*"мочь" + 0.015*"свой" + 0.012*"диалог" + 0.011*"продолжить" + 0.009*"новый" + 0.008*"хотеть" + 0.007*"помочь" + 0.007*"жизнь" + 0.006*"девушка" + 0.006*"думать"
Тема 7	0.013*"мочь" + 0.012*"продолжить" + 0.012*"диалог" + 0.012*"вопрос" + 0.011*"человек" + 0.011*"любить" + 0.011*"парень" + 0.010*"разный" + 0.009*"помогать" + 0.009*"умный"
Тема 8	0.027*"игра" + 0.013*"семья" + 0.011*"играть" + 0.010*"свой" + 0.009*"любить" + 0.008*"время" + 0.008*"друг" + 0.008*"продолжить" + 0.008*"диалог" + 0.007*"настольный"
Тема 9	0.054*"любить" + 0.017*"работать" + 0.015*"диалог" + 0.015*"продолжить" + 0.011*"ребёнок" + 0.010*"заниматься" + 0.009*"хотеть" + 0.009*"жить" + 0.009*"собака" + 0.009*"работа"
Тема 10	0.015*"мочь" + 0.013*"диалог" + 0.012*"продолжить" + 0.012*"вопрос" + 0.011*"разный" + 0.010*"парень" + 0.010*"помогать" + 0.010*"любить" + 0.009*"умный" + 0.009*"консультант"

Проведенный тематический анализ методом латентного размещения Дирихле (LDA) позволяет предположить, что в исследуемом корпусе диалогов присутствуют устойчивые семантические паттерны, проявляющиеся через распределение лексем с наибольшими весовыми коэффициентами. Наблюдается высокая частотность метадискурсивных маркеров («диалог», «продолжить»), что свидетельствует о преобладании рефлексивных речевых актов, направленных на поддержание коммуникации, в то время как лексемы «любить», «нравиться», «хотеть», «свой», маркируют зону субъективных оценочных высказываний. Тематическая структура выявленных тем демонстрирует тенденцию к дифференциации на несколько концептуальных групп: досуговая тематика (кинематограф, музыкальная культура, игры и хобби), когнитивно-прагматическая тематика (консультативные речевые жанры, предполагающие апелляцию к экспертному знанию или оценке), социоцентрическая (межличностные отношения, семейная коммуникация, психология отношений) и аксиологическая тематика (система личностных предпочтений и ценностных ориентаций, вопросы саморазвития). Наличие пересекающихся тематических кластеров (например, тематические кластеры 2 и 10) указывает на необходимость дальнейшей оптимизации параметров модели для минимизации семантического перекрытия.

Несмотря на полученные выводы, позволяющие прогнозировать тематическое распределение диалогов в корпусе, в ходе автоматического определения тематики диалогов с применением метода LDA возникают трудности, связанные с небольшим объёмом текстов и неформальным характером большинства из них. Предварительная процедура очистки диалогов от стоп-слов (предлогов, союзов, частиц) не позволяет избавиться от малоинформативных слов («мочь», «знать», «хотеть», «время» и проч.), которые препятствуют адекватной оценке тематической направленности диалогов.

В качестве следующего метода был применен подход с использованием предобученной большой языковой модели Bert, которая позволяет учитывать контекстуальные семантические связи слов и, следовательно, более эффективно

справляется с задачей тематического представления. В ходе реализации модели были получены 10 тематических кластеров, представленных в Таблице 11.

Таблица 11. Тематическая классификация диалогов на основе модели Bert

№ п/п	Кластер	Примеры (фрагменты)
1.	Кластер 0	Собеседник: Ты знаешь, я заметил, что в последнее время стиль в кино стал играть все большую роль... Собеседник: Я всегда была увлечена цитатами и мудрыми высказываниями...
2.	Кластер 1	Собеседник: Как подать на развод? .. Собеседник: привет. Ты: Привет. Расскажи о себе, давай познакомимся...
3.	Кластер 2	Собеседник: Как составить программу домашнего обучения?.. Собеседник: Как использовать чужие видео на YouTube без проблем?..
4.	Кластер 3	Собеседник: Привет Как дела?.. Собеседник: Чем бы мне сегодня заняться?..
5.	Кластер 4	Собеседник: Ты знаешь, я никогда не могу найти долгосрочные отношения... Собеседник: Понятие зрелости – В контексте личности...
6.	Кластер 5	Собеседник: Меня часто тошнит от машины... Собеседник: Я ипохондрик, и это медленно разрушает мое психическое здоровье...
7.	Кластер 6	Собеседник: Что мне нужно сделать, чтобы стать сертифицированным агентом по недвижимости?.. Собеседник: Как найти своего ангела-хранителя по имени только?..
8.	Кластер 7	Собеседник: Привет. Я пенсионер. Уже не работаю... Собеседник: Здравствуйте рад новому знакомству...

9.	Кластер 8	Собеседник: Привет. Как дела... Собеседник: Я люблю шить подумываю уже шить на заказ...
10.	Кластер 9	Собеседник: Я хочу купить дом и участок через брокера... Собеседник: Найди сумму чисел в предложении...

Анализ диалогов, включенных в разные кластеры, показывает наличие тематических пересечений, не учтенных при автоматическом тематическом моделировании. Так, группа диалогов, связанная с формальным запросом информации, была разбита на различные кластеры исходя из тематического содержания запроса без учета структуры диалога и его коммуникативных особенностей (кластеры 1, 2, 6, 9). В то же время в один кластер были объединены диалоги, имеющие различную тематическую направленность (кластеры 1, 6).

Уменьшение размерности векторов эмбедингов с их последующей визуализацией в двухмерном пространстве позволяет подтвердить избыточность выделенных кластеров (рис. 19). Визуальное представление расстояний между эмбедингами слов позволяет выделить 5 кластеров, среди которых один – наиболее крупный, остальные представляют собой достаточно обособленные и незначительные по объему группы. Особняком стоит кластер, представленный на рис. 19 в правом верхнем углу: вытянутый вдоль оси Y, он неоднороден в тематическом отношении, однако демонстрирует значительное тематическое сходство входящих в него диалогов.

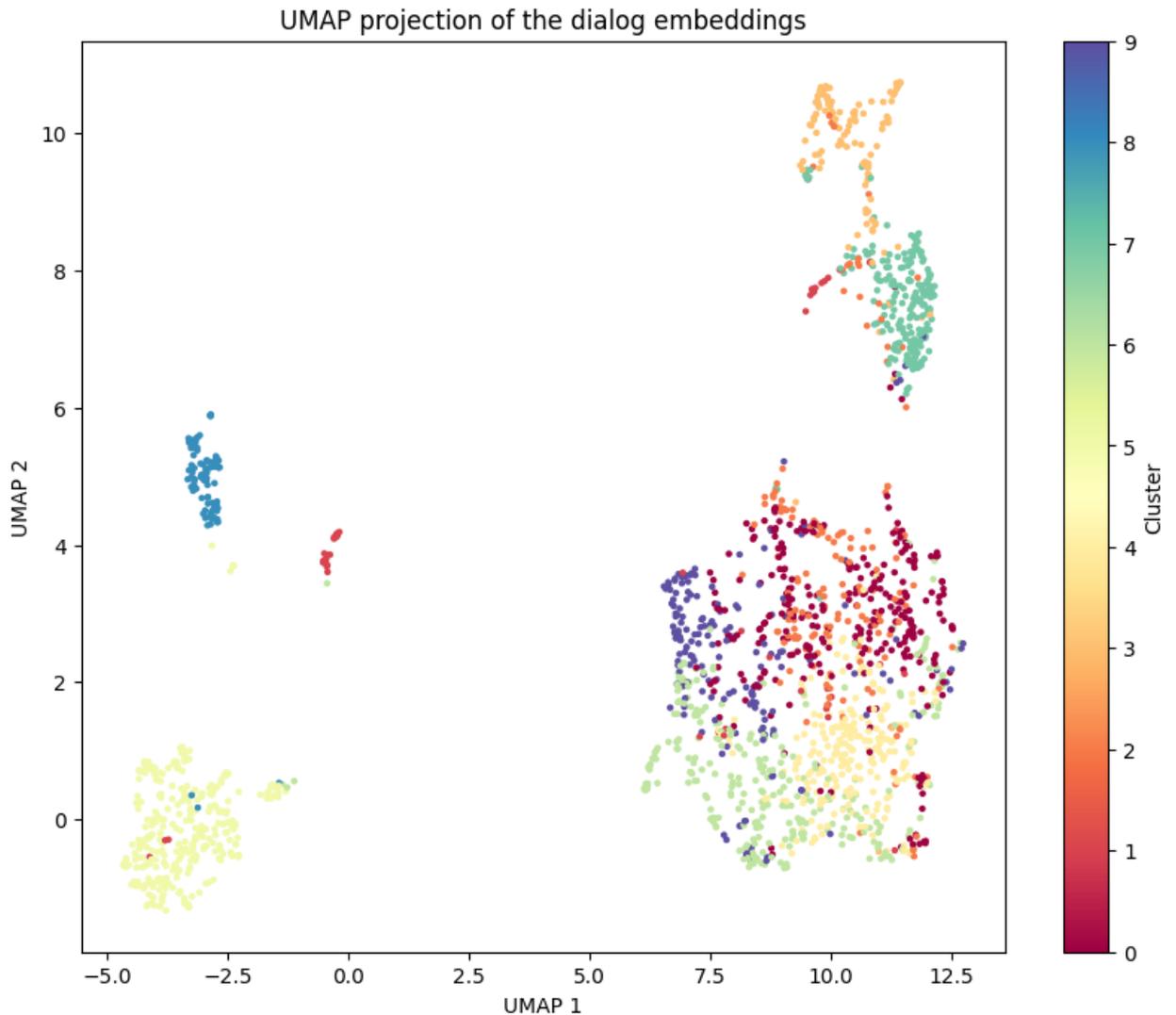


Рисунок 19. Визуализация векторов эмбедингов в двухмерном пространстве признаков

Результаты тематического моделирования с использованием LDA и модели Bert были сопоставлены с данными, полученными в ходе машинного обучения с использованием обучающей выборки диалогов, размеченных вручную. Темы, выделенные в обучающей выборке, были распределены моделью по диалогам неравномерно.

Тематическое поле большинства диалогов (свыше 80%) – ситуация знакомства, обсуждение бытовых планов и дел, домашних питомцев. Небольшой процент диалогов составляют специфические диалоги-запросы информации, выделяющиеся на фоне фатической коммуникативной модели остальных диалоговых текстов направленностью на получение определённой информации.

Такие диалоги наиболее соответствуют тематическим запросам пользователя к диалоговым системам при поиске информации, следовательно, являются формальными и не воспроизводят характерные паттерны неформальной коммуникации.

В связи с ограниченной возможностью автоматического определения тематики диалогов результаты были скорректированы вручную. Итоговое распределение диалогов из выбранного подмножества по тематическим классам в процентах от общего количества представлено на рис. 20.

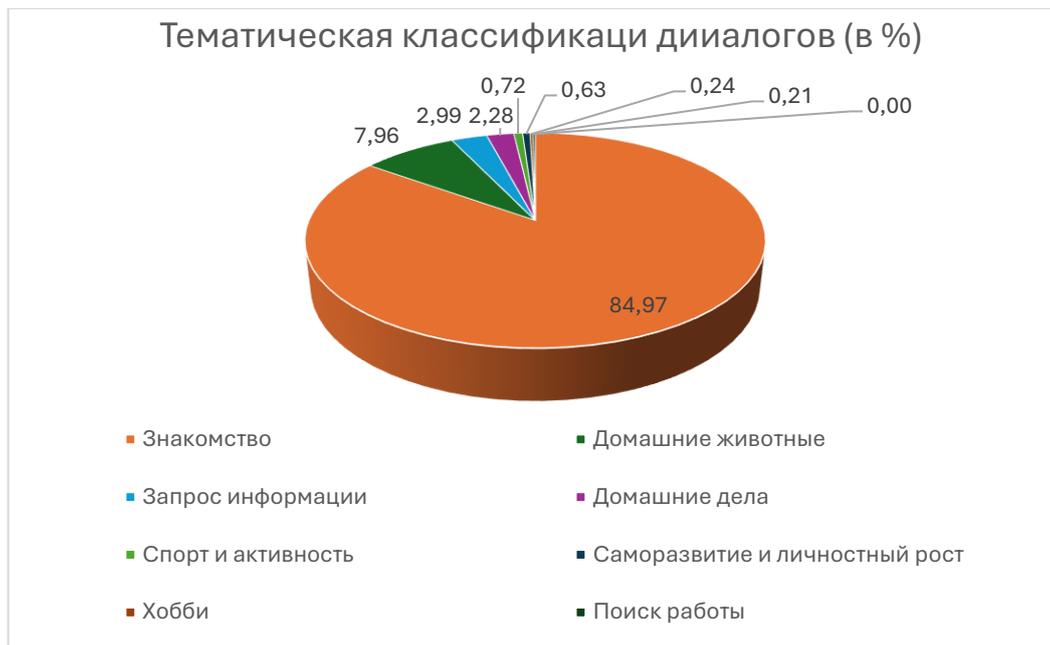


Рисунок 20. Распределение тем в диалогах (ручной анализ)

Для автоматизации процесса разметки реплик использовались регулярные выражения, возможность применения которых обусловлена формализованной структурой исходных текстов выбранного подмножества. Каждый диалог в представленном наборе данных имеет однотипную структуру: инициирует диалог реплика персонажа «Собеседник»; далее реплики персонажей «Ты» и «Собеседник» последовательно чередуются, завершаясь всегда репликой персонажа «Ты».

Это структурное отличие сгенерированного диалога от естественной коммуникации задано намерено в процессе генерации, отражает сценарную природу диалогового взаимодействия в процессе генерации текста, но при этом

формальная предсказуемость структуры позволяет эффективно использовать регулярные выражения для извлечения необходимой информации.

В результате применения регулярного выражения к разметке реплик были извлечены именованные группы, захватывающие номинацию персонажа, а также реплики данного персонажа, что позволяет создать корпус реплик каждой нейросетевой маски для их дальнейшего изучения.

3.3. Структурно-композиционные особенности сгенерированных диалоговых текстов

Для изучения структуры диалогов был создан корпус текстов с применением библиотеки Gensim на языке программирования Python. Корпус основан на структурированных текстовых данных, содержащих исследуемое подмножество диалогов, снабженных полями:

- `speaker` – персонаж, которому принадлежит реплика в диалоге;
- `utterance` – исходный текст реплики;
- `processed_text` – предварительно обработанный текст реплики, подвергнутый процедуре приведения к нижнему регистру, удаления стоп-слов и токенизации;
- `predicted_topic` – тема диалога.

После предварительной обработки реплик диалога каждый элемент поля добавляется в корпус в виде словаря, который загружается в DataFrame с помощью библиотеки Pandas для представления данных в табличном виде.

С помощью полученного корпуса были проанализированы формальные и статистические характеристики реплик в разбивке по спикерам и темам:

- средняя длина реплик;
- средняя длина предложений в составе реплик;
- индекс лексического разнообразия реплик;
- индекс Флеша;
- доля вопросительных предложений;

- доля восклицательных предложений;
- доля повествовательных предложений.

Полученные результаты представлены в Таблице 12.

Таблица 12. Формальные характеристики диалогов

Спикер	Тема	Средняя длина реплик	Средняя длина предложений	Индекс лексического разнообразия	Индекс Флеша	Доля вопросительных предложений	Доля восклицательных предложений	Доля повествовательных предложений
Собеседник	Домашние дела	4,35	1,96	0,61	117,82	51,22	13,82	34,96
Собеседник	Домашние животные	8,30	2,17	0,56	113,81	46,93	24,53	28,53
Собеседник	Запрос информации	10,62	1,97	0,65	111,45	68,42	22,11	9,47
Собеседник	Знакомство	5,31	1,76	0,34	116,84	52,69	11,06	36,25
Собеседник	Поиск работы	13,17	2,00	0,81	108,87	50,00	0,00	50,00
Собеседник	Саморазвитие и личностный рост	8,58	2,04	0,61	113,52	83,33	29,17	0,00
Собеседник	Спорт и активность	7,55	2,26	0,68	114,57	33,96	24,53	41,51
Собеседник	Хобби	1,86	3,36	0,54	120,35	0,00	7,14	92,86
Ты	Домашние дела	13,43	4,04	0,55	108,60	33,55	19,74	46,71
Ты	Домашние животные	51,68	6,37	0,28	69,78	19,19	24,74	56,07
Ты	Запрос информации	97,93	3,07	0,27	22,84	14,47	23,51	62,02

Спикер	Тема	Средняя длина реплик	Средняя длина предложений	Индекс лексического разнообразия	Индекс Флеша	Доля вопросительных предложений	Доля восклицательных предложений	Доля повествовательных предложений
Ты	Знакомство	23,48	4,95	0,19	98,40	22,12	15,38	62,50
Ты	Поиск работы	55,18	3,39	0,61	66,23	11,76	0,00	88,24
Ты	Саморазвитие и личностный рост	28,03	3,43	0,61	93,78	15,52	39,66	44,83
Ты	Спорт и активность	37,52	2,59	0,56	84,15	28,57	14,29	57,14
Ты	Хобби	7,24	92,94	0,70	114,89	11,76	23,53	64,71

Анализируя представленные статистические данные, полученные на основе автоматического подсчета метрик, можно отметить следующее:

1. У персонажа «Ты» средняя длина реплик, а также средняя длина предложений (особенно в темах, требующих развернутого ответа) больше, чем у персонажа «Собеседник», так как диалоги основаны на продуцировании содержательных ответов от лица нейросетевой маски.

2. Показатель средней длины реплики наиболее вариативен в разрезе темы диалога и выше в тех случаях, где тематика беседы касается сообщения сведений (запрос информации), описания (поиск работы, домашние животные, спорт и активность).

3. Индекс Флеша, отражающий показатель читаемости текста, варьируется от 60 до 120, находясь в диапазоне текста нормального уровня сложности, что также объясняется установкой на воспроизведение неформальной манеры

общения, исключая сложнейшие профессиональные или специальные научные темы.

4. В репликах персонажа «Собеседник» чаще используются вопросительные предложения, так как его реплики инициируют каждый диалог. У персонажа «Ты» доля вопросительных предложений значительно ниже, поскольку ему отводится реактивная роль в обмене репликами.

5. Доля восклицательных предложений в среднем одинакова для обоих персонажей и остается незначительной, но выше для тех тем, которые связаны с обсуждением лично значимых вопросов (домашние животные, саморазвитие и личностный рост, хобби).

6. Доля повествовательных предложений выше для персонажа «Ты», что связано, во-первых, со структурой диалога (реплики персонажа «Ты» представляет собой, как правило, ответы на вопросы), а во-вторых, с функцией персонажа «Ты» в модели генерации диалоговых текстов, где ему отводится роль информанта, сообщающего сведения по запросу, дающего разъяснения, комментирующего реплики собеседника.

Таким образом, метрики, полученные в результате автоматического анализа диалогов, показывают зависимость формальной структуры генерируемых реплик от обсуждаемой темы и спикера, которому они принадлежат. Структура сгенерированных текстов указывает на однотипность диалогового взаимодействия между участниками, инициированного преимущественно вопросительными неразвернутыми репликами персонажа «Собеседник» и поддерживаемого ответными репликами персонажа «Ты», которые носят более развернутый и информативный характер. Такая асимметричность в распределении формальных характеристик реплик, воспроизводящихся в различных тематических блоках, указывает на шаблонность, стереотипность структуры исследуемых текстов.

Другой важной характеристикой, отражающей структуру диалога, являются переходы между репликами, позволяющие определить паттерны при построении диалогового взаимодействия. С целью изучения переходов между репликами в диалогах были автоматически выявлены типы переходов между репликами с

подсчетом частоты их воспроизводимости. Результаты для наиболее часто встречающихся типов переходов между репликами отражены в Таблице 13.

Таблица 13. Частота переходов между репликами спикеров в диалогах (на примере 20 наиболее частотных)

Индекс	Переход	Частота
0	Привет. -> Привет.	49
1	привет. -> Привет.	10
2	Привет. -> Как дела?	9
3	Привет! -> Привет.	9
4	привет -> здравствуй	5
5	Привет. -> Здравствуй.	4
6	Привет. -> Чем занимаешься?	4
7	Привет. -> привет.	4
8	Привет. -> Привет!	4
9	Привет. -> Расскажи о себе.	3
10	Привет. -> Привет, как дела?	3
11	Привет. -> Приветик.	3
12	привет. -> Как дела?	3
13	Пока -> Привет.	3
14	Приятно познакомиться. -> Взаимно.	3
15	Привет. Вас также. -> Спасибо. Блинчики вот пеку. Люблю блинчики. Особенно с начинками. Давай на ты. А то чувствую себя каким-то пенсионером.	2
16	Ужасно! А есть еще какие-то интересные факты о Нероне? -> Да, например, его называли певцом, писателем и владыкой мира из-за его увлечения музыкой и литературой. Он даже устраивал концерты и сам играл на фортепиано.	2
17	Да, например, его называли певцом, писателем и владыкой мира из-за его увлечения музыкой и литературой. Он даже устраивал концерты и сам играл на фортепиано. -> Как необычно! А еще у него была интересная особенность - он вел себя необычно во время пожара в Риме, правда?	2
18	Как необычно! А еще у него была интересная особенность - он вел себя необычно во время пожара в Риме, правда? -> Да, это верно. Во время пожара, который уничтожил большую часть Рима, Нерон, сидя на	2

Индекс	Переход	Частота
	колоннаде, играл на флейте и пел песню о городе, разрушение которого наблюдал.	
19	Привет. -> Как тебя зовут?	2

Анализ данных показывает, что наиболее частым переходом между репликами является взаимное приветствие (в различных вариантах оно встречается с частотой от 4 до 49), менее частотными (частота – 2 и менее) являются ответы на вопрос («А есть еще какие-то интересные факты о Нероне? -> Да, например, его называли певцом, писателем и владыкой мира...») или оценочная реплика («Он даже устраивал концерты и сам играл на фортепиано. -> Как необычно!..»).

Для исследования структуры диалогов по различным темам целесообразно классифицировать переходы между репликами, выделив несколько наиболее часто встречающихся типов: приветствия, запросы информации, комментарии, ответы на вопросы, сообщение информации, эмоционально-оценочные реплики. Примеры основных типов переходов приведены в Таблице 14.

Таблица 14. Примеры типов переходов между репликами

Тип перехода	Пример реплик из диалога
Приветствие	Привет. -> Привет.
Запросы информации	Очень люблю книги, а как твое мнение о книгах? -> Нормально, но больше люблю слушать музыку.
Комментарии	Да, он большой любитель пиццы. -> Я кстати пока не встретил свою половинку- нет у меня такой Марина.
Ответ на вопрос	Как ты относишься к экологии? -> Я считаю, что охрана природы и экологическое сознание - очень важные вопросы.
Сообщение информации	Ты работаешь? -> Мм, пока нет, но надеюсь, что пройду кастинг на фотомодель
Эмоционально-оценочная реплика	Мейкун. -> Оооо, они просто шикарные.

Тип перехода	Пример реплик из диалога
Этикетные реплики	Приятно познакомиться. -> Взаимно.

Распределение реплик по типам переходов предоставлено на рис. 21.



Рисунок 21. Распределение реплик в диалогах по типам переходов между репликами (в %).

Полученные данные свидетельствуют о том, что большая часть переходов между репликами относится к запросу информации (50,70%), а также сообщению информации (22,17%) или ответу на вопрос (14,11%). В целом такая структура коррелирует с данными, полученными в ходе автоматического расчета доли вопросительных и повествовательных предложений в диалогах, и показывает, что основная модель диалогового взаимодействия – вопросно-ответная.

Исследование рекуррентных паттернов переходов между репликами в разрезе темы и спикера позволяет установить, что переходы между репликами наиболее частотны для обширной темы «Знакомство», максимально соответствующей естественной диалогической речи и включающей почти 80% от общего числа диалогов (рис. 22).

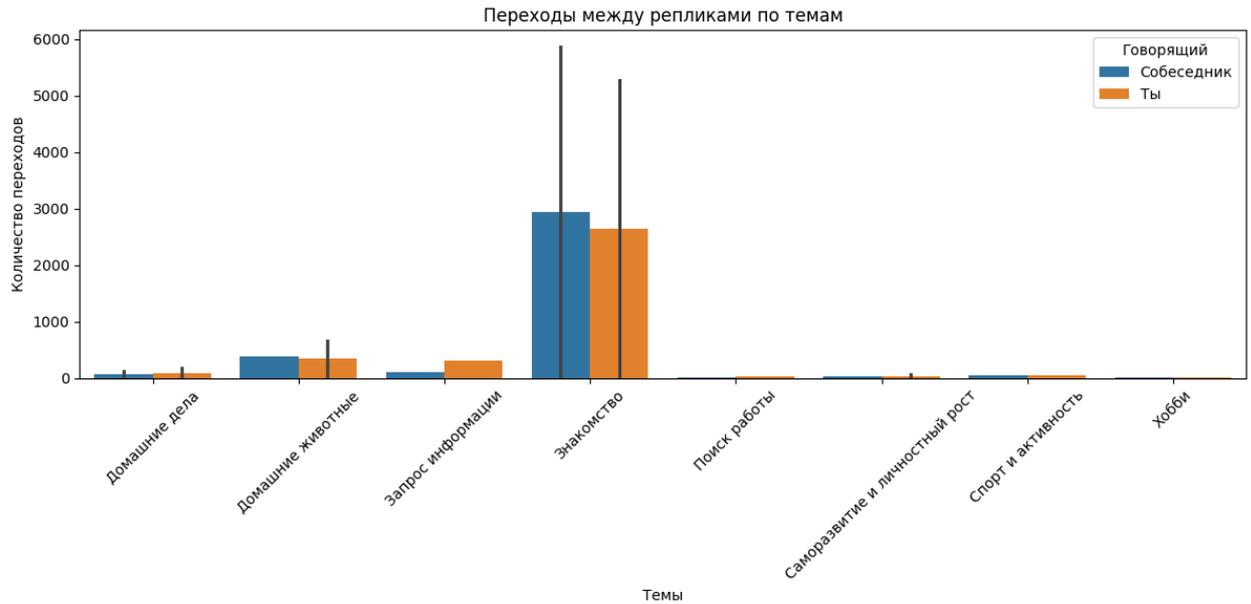


Рисунок 22. Анализ типов переходов (речевых актов) между репликами диалогов

Для остальных тем такие переходы по частотности остаются незначительными, что свидетельствует о малом обмене репликами между персонажами диалога и приближении вопросно-ответного взаимодействия к развернутым монологическим ответам нейросетевой модели (персонаж «Ты») на запрос собеседника. Рекуррентные паттерны переходов между репликами в разрезе темы и типа перехода также демонстрируют наибольшее разнообразие категорий переходов между репликами в теме «Знакомства», где встречаются в разных соотношениях все типы переходов: приветствие, запрос информации, ответ на вопрос, комментарий, эмоционально-оценочные реплики, сообщение информации (рис. 23). Запросы информации, как и сообщение информации, тоже доминируют над остальными типами переходов, что говорит о преимущественно вопросно-ответной коммуникативной модели взаимодействия.

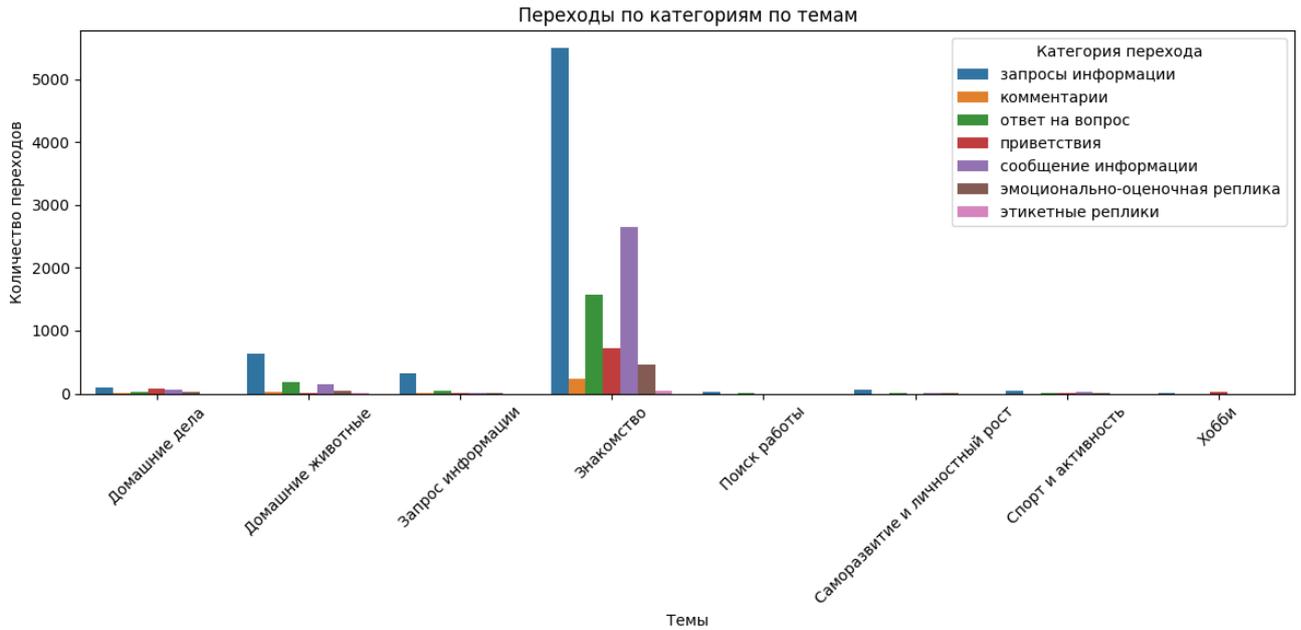


Рисунок 23. Анализ типов переходов (речевых актов) между репликами диалогов в разрезе обсуждаемых тем

Статистические параметры, характеризующие переходы между репликами, свидетельствует о формальной рекуррентной структуре диалога, в котором каждая реплика непосредственно связана с предыдущей. Эта структура воспроизводит модель взаимодействия спикеров, основанную на использовании устойчивых схем преобразования реплик с варьированием переходов между репликами от простых «этикетных» приветствий и контактоустанавливающих фраз до более сложных реакций (ответы на вопросы, сообщение сведений, уточняющие вопросы, оценочные реплики). Трансформационный анализ описывает механизм формирования более сложных языковых конструкций, соответствующих поверхностным структурам диалога, на основе глубинных структур – первичных элементов взаимодействия.

Структурная модель сгенерированного диалогового текста включает регулярные паттерны развертывания коммуникативного взаимодействия, которое начинается с простых форм приветствия и постепенно усложняется по мере углубления темы. В символической нотации данная структурная схема

развертывания диалогового текста может быть представлена следующим образом (рис. 24).

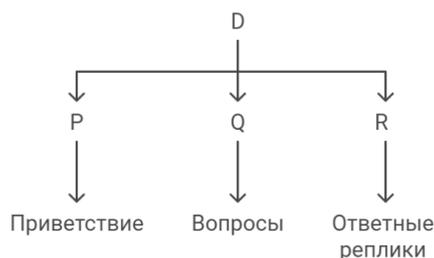


Рисунок 24. Модель развертывания диалога в системе текстогенерации

D представляет собой основной узел диалогового взаимодействия, который включает элементы глубинной структуры: P (приветствие), Q (вопросы), R (ответные реплики – углубление обсуждения), а также их поверхностные репрезентации в виде конкретного языкового наполнения реплик. Преобразование глубинных структур в поверхностные происходит по правилам, заданным алгоритмом машинного обучения, вне связи с контекстом и ситуацией общения, как это происходит в естественной диалоговой коммуникации.

3.4. Морфолого-синтаксическая организация сгенерированных диалогов

Изучение морфолого-синтаксической структуры сгенерированных диалоговых текстов позволяет выявить устойчивые паттерны формальной организации текста, проявляющиеся в частотности словоупотребления и использования синтаксических конструкций, определяющих грамматику сгенерированного текста.

Количественный анализ диалогов, основанный на подсчете распределения слов разных частей речи в процентном отношении к общему числу всех лексем, представлен на рис. 25.

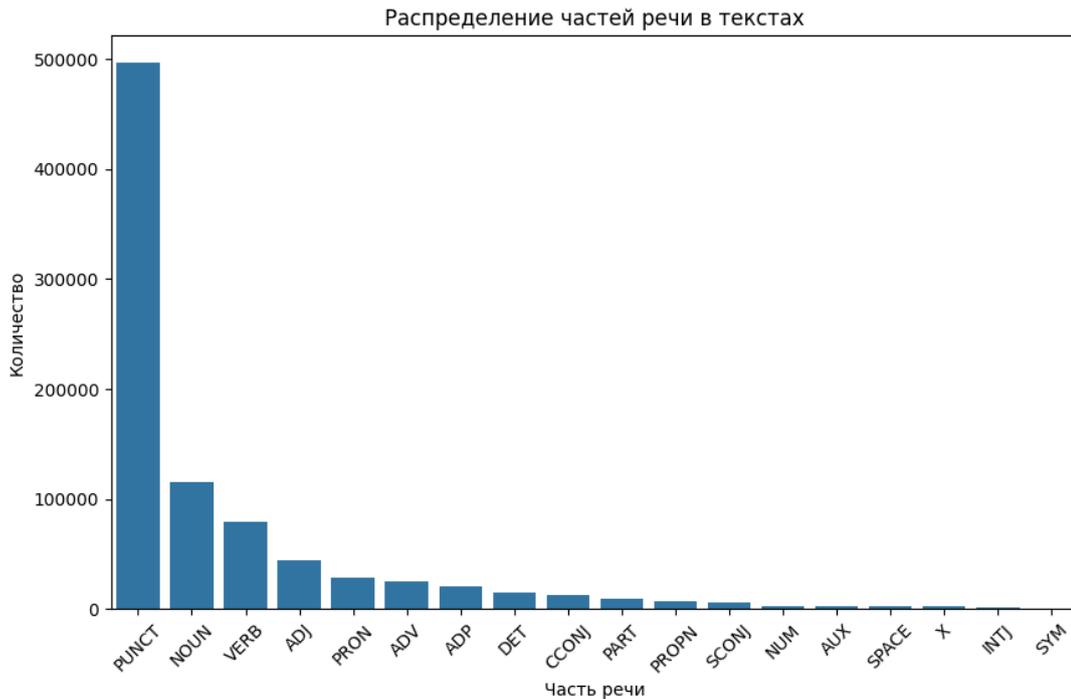


Рисунок 25. Распределение слов разных частей речи в сгенерированных диалогах

Анализ данных полученных данных позволяет выделить формальные характеристики, характерные для морфологической организации диалогов:

высокая частотность субстантивной лексики по сравнению со словами других частей речи указывает на фактологический или описательный характер диалогов, которые связаны с обсуждением тем, ситуаций и событий вне временного и пространственного контекста их развертывания;

меньший процент глагольной лексики (глаголов меньше, чем имен существительных почти в 1,4 раза) свидетельствует о том, что тематика диалогов слабо связана с динамическим аспектом обсуждаемой коммуникативной ситуации и персональным опытом спикеров, который не маркируется в речи соответствующей глагольной лексикой, характерной для нарратива воспоминаний, рассказа о событии, случая из жизни и проч.;

индекс местоименности (который вычисляется как отношение количества местоимений к общему числу полнозначных лексем) принимает значения 0,037, что является в большей степени характерным для книжных формальных стилей (научного и официально-делового), в то время как в разговорном стиле он

значительно выше в силу особой роли дейктических элементов в ситуации общения, компоненты которой ситуативно обусловлены;

значимым показателем является низкое процентное содержание междометий (примерно 0,1%), указывающее на невысокую экспрессивность диалогического взаимодействия.

Данные показатели могут быть дополнены анализом конкретных лексико-семантических групп, связанных с определёнными формально-грамматическими зависимостями. Тем не менее формальный количественный анализ показывает, что морфологическая структура сгенерированных текстов отличается от наблюдаемых в устной и письменной естественной диалогической коммуникации закономерностей. В сгенерированных диалогах на первый план выходит не фатическая составляющая общения с присущей ей экспрессией и оценочностью, а информационная, связанная с генерацией и передачей фактических сведений, релевантных содержанию запроса на генерацию.

Слабая выраженность лексико-морфологических признаков «биографического нарратива», связанных с частотными для естественных диалогов ситуациями воспоминаний, рассказов о случаях из жизни, пересказа чужой речи, обстоятельств встреч, бесед, сопровождающихся использованием экспрессивно окрашенных языковых средств, маркирует автономность сгенерированных диалогов от «опыта» спикера, ограничивая диапазон речевых реакций заданными промптом сценариями.

Анализ формальных показателей синтаксических зависимостей в сгенерированных диалогах позволяет выявить следующие особенности:

избыточность пунктуации, как следствие, правильность синтаксической организации предложений, характерная для сгенерированных текстов и часто нарушаемая в текстах естественной письменной речи;

значимая доля корневых зависимостей (ROOT) указывает на большой процент синтаксически независимых предложений, более характерный для монологической речи, чем для диалога, в котором реплики носят реактивный характер и часто характеризуются неполнотой синтаксической структуры (в

диалогах преобладают эллиптические конструкции, ситуативно и контекстуально неполные предложения);

высокий показатель case (19444) отражает наличие разнообразных предложно-падежных конструкций, свидетельствуя о достаточно сложной грамматической организации большинства диалогов;

значения показателей зависимостей advmod (17197) и amod (14165) указывают на обстоятельственные и атрибутивные определения, сопутствующие глагольной и субстантивной лексике и придающие содержанию диалогов описательный характер;

зависимость obj (12836), указывающая на количество конструкций с прямым дополнением, позволяет сделать вывод о преобладании в сгенерированных диалогах прямого порядка слов, который не является характерным для естественной коммуникации.

Распределение показателей проанализированных синтаксических зависимостей в диалогических сгенерированных текстах приведено на рис. 26.

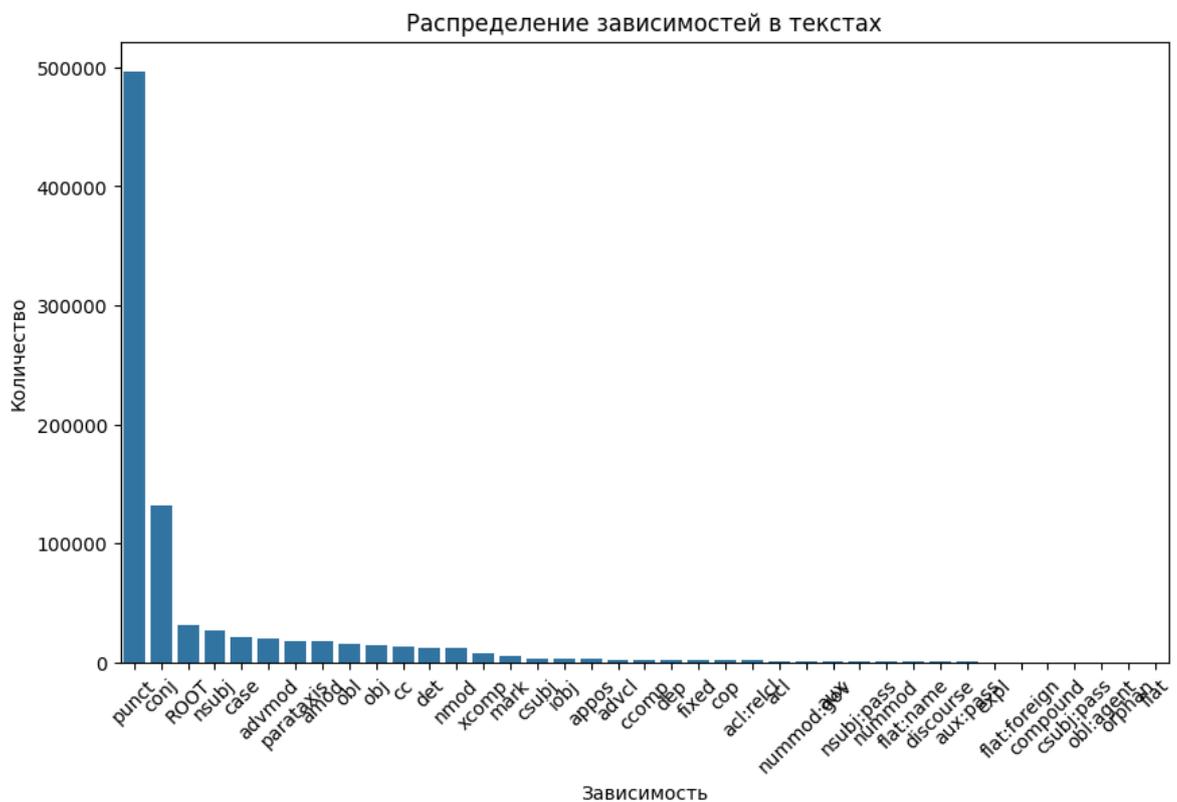


Рисунок 26. Распределение синтаксических зависимостей в сгенерированных диалоговых текстах

Выявленные паттерны свидетельствуют о хорошей структурированности сгенерированных диалогов, которые тяготеют к описательности, статичности, преобладанию конкретной фактологической и информационной составляющей над ситуативно-динамическими компонентами, заданными контекстом. Тяготение формально-грамматической организации сгенерированных диалогов к письменной монологической коммуникации обуславливает их грамматическую правильность, разнообразие используемых типов конструкций, соблюдение пунктуационных правил. Вместе с тем преобладание субстантивной лексики, адвербиальных и именных атрибутивов над глагольной, местоименной и междометной лексикой указывает на специфику грамматической природы сгенерированного текста, у которого ослаблены событийная, экспрессивно-речевая и ситуативная информационная составляющие.

Важной составляющей сгенерированного текста также являются ошибки, которые могут быть классифицированы в соответствии с принятой в языкознании типологией нарушения языковых норм. В изучаемых текстах наблюдаются следующие виды ошибок:

орфографические, связанные с искажением графического облика слов (ошибки, похожие на опечатки), но, как правило, не отражающие нарушение норм правописания (например: «Я активно читаю книги, люблю погулять в парке и *покуцать* пиццу с друзьями», диалог 70; «привет я не работаю я *люблб* петь и готовить», диалог 94);

пунктуационные, связанные в основном с отсутствием знаков препинания при вводных словах (например: «Потом *правда* тяжело на работу возвращаться», диалог 3) или отражающие тенденцию к гиперкоррекции за счёт постановки лишних знаков препинания (например: «*Также,* мы постоянно следим за их здоровьем и водим на регулярные ветеринарные осмотры», диалог 7; «Уверена, что с интересной компанией, даже плохая погода становится лучше», диалог 45; «*Однако,* есть несколько вещей, которые можно сделать, чтобы снизить расходы на ремонт и продлить срок службы вашего автомобиля», диалог 74);

грамматические, отражающие нарушение морфолого-синтаксических норм (например: «*Городскую суету исторических европейских городов или отдых на пляже тропического острова – все меня привлекает*», диалог 12);

речевые, возникающие как следствие нарушения норм лексической сочетаемости, употребления слов в значениях, которые в целом несвойственны речевой практике («Также может быть полезно обсудить свой *опыт расизма*, так как это может помочь вам осмыслить то, что произошло, и справиться с эмоциональной *травмой, которую вы пережили*», диалог 47; «в последние годы спор о универсалиях *стал менее центральным* в современной философской дискуссии», диалог 171; «Но как понять, что это именно тот человек? Может быть, я слишком *выбирающий?*», диалог 185).

Выявленные ошибки обусловлены не стратегией алгоритма к имитации естественной письменной речи, а особенностями процесса генерации текста, которые приводят к появлению специфических «сбоев», проявляющихся в виде перестановок и замен букв, генерации лишних символов, аграмматизмов и других признаков речевых аномалий. Сгенерированный текст выглядит более правильным и грамотным по сравнению с естественной письменной речью; ошибки, которые в нем возникают, не имеют системного характера, никак не маркируют особенности идиостиля спикеров, поэтому их частота и типы не могут быть интерпретированы как характерный паттерн сгенерированных диалогических текстов.

3.5. Сравнительный анализ сгенерированных и естественных диалогов

Компоненты лингво-информационной модели искусственного текста диалоговой формы, включающие формально-структурные, содержательные, языковые и коммуникативно-прагматические особенности, требуют сопоставления с компонентами лингвистической модели естественных диалоговых текстов для выявления специфических особенностей процессов текстогенерации.

В качестве материала исследования использованы русскоязычные диалогические тексты устной непубличной коммуникации, размещенные в Национальном корпусе русского языка и агрегированные по следующим параметрам: устная непубличная речь, атрибут авторства – смешанный гендерный признак (женский и мужской). В отношении данных текстов были проведен предварительный количественный и качественный анализ для выявления их пригодности для сравнительного анализа со сгенерированными текстами.

Количественные параметры текстов подкорпуса соответствуют целям и методологии исследования. Объем подкорпуса составляет 1940 диалогических текстов, включающих не менее 10 реплик каждый, что сопоставимо с объемом исследуемых сгенерированных текстов. Сфера функционирования исследуемых текстов (рис. 27) охватывает устную непубличную коммуникацию, что в целом соответствует коммуникативному профилю сгенерированных диалогов.

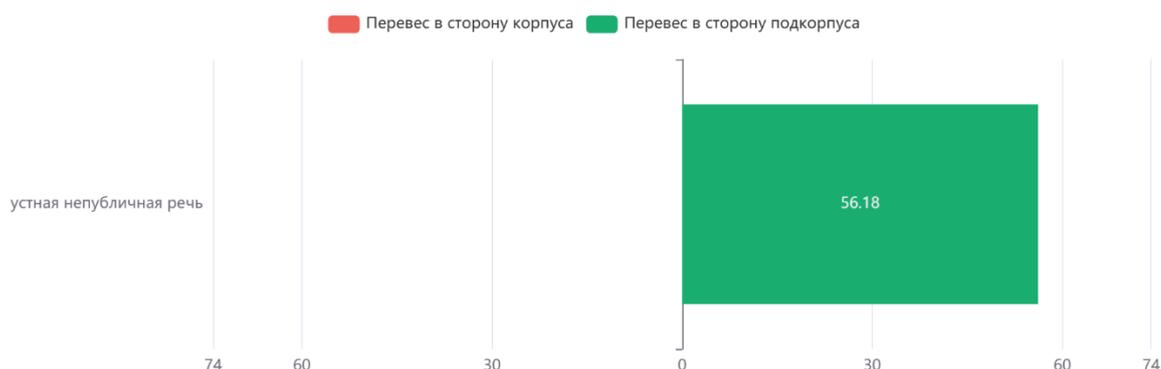


Рисунок 27. Сфера функционирования естественных диалогов в подкорпусе НКРЯ

Качественный анализ естественных диалогов в подкорпусе показывает, что ввиду устной формы их бытования следует признать их ограниченно пригодными для сопоставления с синтетическими диалоговыми текстами, которые создавались как имитация естественной письменной речи. В целях обеспечения объективности сравнительного анализа было решено не учитывать специфические особенности устных диалоговых текстов, включающие акцентологический, просодический,

мелодический рисунок фразы, и сосредоточиться на их тематическом составе, формальных, структурных и грамматических характеристиках.

С целью автоматизации процедуры лингвистического моделирования естественных диалогов был осуществлен анализ текстов с использованием методов корпусной лингвистики, реализованных в функционале ресурса «Национальный корпус русского языка» (анализ контекстного окружения языковых единиц с учетом их семантики). Полученные результаты были дополнены методами качественного анализа. В ходе исследования были установлены:

тематические группы диалоговых сообщений естественной коммуникации;
формально-структурные особенности естественных диалогов;
особенности морфолого-синтаксического строения естественных диалогов.

Тематические группы сгенерированных сообщений представлены в Таблице 15.

Таблица 15. Тематические кластеры естественных диалогов

Тема	Количество диалогов	Доля от общего количества в %
Частная жизнь	1704	87,84
Наука и технологии	26	1,34
Искусство, культура и искусствоведение	68	3,50
Образование	41	2,11
Политика и общественная жизнь	48	2,47
Философия и религия	11	0,57
Здоровье и медицина	27	1,39
Досуг, зрелища и развлечения	31	1,60
Спорт	9	0,46

Тема	Количество диалогов	Доля от общего количества в %
Путешествия	11	0,57
Бизнес, экономика и финансы	82	4,25
История	3	0,15
Техника и промышленность	13	0,67
Криминал и право	6	0,31
Дом и домашнее хозяйство	6	0,31
Информатика, астрология и эзотерика	3	0,15
Вымышленный мир	4	0,21
Организация и управление	2	0,1
Финансы	1	0,05

Визуализация тематических кластеров в процентном отношении от общего числа диалогов представлена на рис. 28.

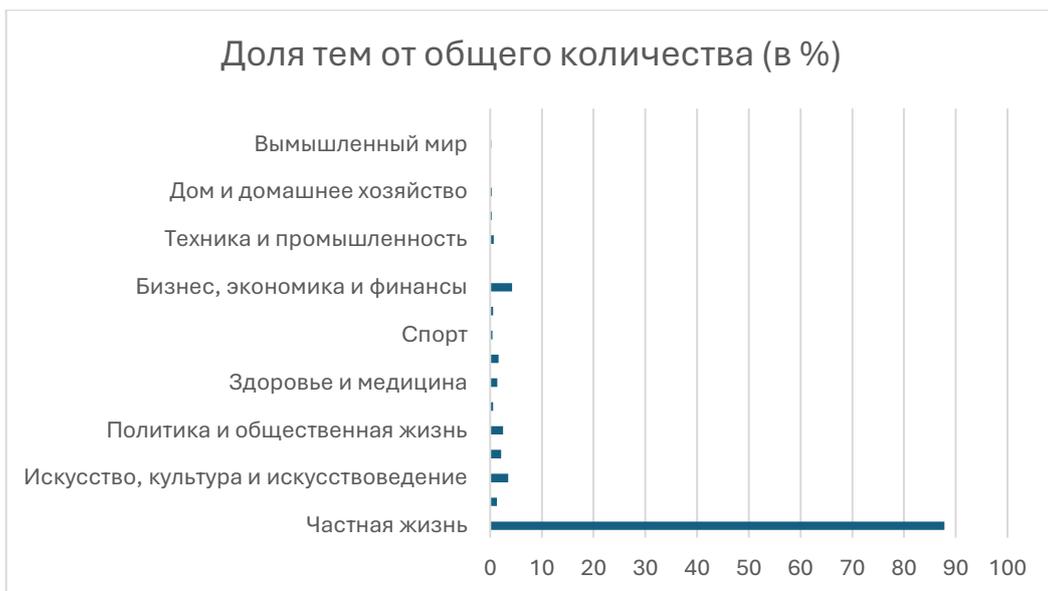


Рисунок 28. Тематическое распределение естественных диалогов

При сопоставлении с тематическим составом сгенерированных диалоговых текстов можно выделить следующие черты сходства и различия:

искусственные диалогические тексты, как и естественные диалоги, ориентированы на простые взаимодействия в контексте непубличной коммуникации, что обуславливает подбор тем, связанных преимущественно с событиями частной жизни (общение в кругу семьи и друзей, знакомство, обсуждение бытовых тем, советы и рекомендации личного характера и проч.);

тематический состав синтетических диалогов преимущественно однотипен и касается определённого круга тематических паттернов, которые принято ассоциировать с неформальным общением (хобби, домашние животные, домашние дела, спорт);

тематика естественных диалогов расширяется за счет привлечения контекста, включающего, помимо ситуации общения, и фоновые знания (опыт коммуникации) участников взаимодействия, включая их эмоционально-оценочные речевые реакции, выраженные в диалогах прагматические установки по отношению к собеседнику;

тематическая ограниченность искусственных диалоговых текстов обуславливает их содержательную лаконичность, в то время как в естественных диалогах наблюдается большая информационная избыточность при обсуждении темы за счет повторов, утоняющих формулировок, маркеров экспрессии и оценочности;

важной составляющей тематики естественных диалоговых текстов являются маркеры принадлежности к определённому историческому контексту, в котором разворачивается беседа, но «приметы времени» в целом практически отсутствуют в сгенерированных текстах. Например, в естественных диалогах упоминаются современные произведения литературы и кинематографа (серия книг о «Гарри Поттере», диалог 11, художественный фильм «Аватар», диалог 26), популярные актеры (Дэниел Крейг, сыгравший Джеймса Бонда, диалог 31), упоминание реалий эпохи пандемии (запрет на посещение магазина без маски, диалог 54) и др.

Наблюдаемые тематические различия в составе диалогических текстов связаны, на наш взгляд, со следующими факторами:

сгенерированные тексты, охватывающие в основном тему знакомства, предназначены для создания начальных точек общения для обучения взаимодействию с пользователем, что предполагает использование форм сценарного взаимодействия в рамках ситуации знакомства, обуславливающих тематическую ограниченность и однотипность сгенерированных диалогов;

большее тематическое разнообразие естественных диалогов за счет погружения в более широкий контекст (наука, искусство, политика) показывает вовлеченность коммуникантов в сложные формы социального взаимодействия, которые опосредованы общением;

доминирование ситуативно обусловленных форм коммуникации в естественной речи, где диалог выступает частью социальных интеракций, связывающих тему беседы с социально-ролевым статусом субъекта (семейным или общественным); в связи с этим в диалогах четко маркируются социально-ролевой статус коммуникантов: члены семьи (дети – родители), члены профессиональных сообществ (преподаватель – обучающийся), участники экономических отношений (покупатель – продавец) и проч.;

преимущественно устный способ коммуникации и, как следствие, больший процент отклонения от темы, наличие признаков спонтанности речи, что отражается и на характере диалогового взаимодействия, который приобретает менее предсказуемый характер и отличается тенденцией к расширению (изменению) первоначальной темы обсуждения.

Структурно-композиционные особенности естественных диалогов обусловлены устной формой их бытования. В целом структура большинства естественных диалогов, представленных в исследуемой выборке, отличается фрагментарностью. Количественный анализ состава реплик показывает, что инициатива в диалоге распределяется равномерно, формальное преобладание реплик одного из коммуникантов наблюдается в жанровом формате интервью¹⁸, в

¹⁸ Например: «Ксения Олеговна А. Расскажи/ пожалуйста/ о своём самом любимом фильме. Дарья Андреевна К. М... мой любимый фильм – это «Аватар» первая и вторые части. Он меня привлекает тем/ что там м... оч красивая графика/ аа которая детально проработана/ что складывается впечатление/ как будто это прям

то время как неформальные беседы (в том числе опосредованные телефонной связью) состоят преимущественно из коротких (4–5 слов) реплик, которыми обмениваются собеседники. Например: «№1. Привет, как дела??? №2. Привет. Дела нормально. Отдыхаю. Погода хорошая. А ты как? № 1. Вот и я отдыхаю, что нового? Как твоя блондинка?» (разговор на улице между женщиной и женщиной, диалог 3); «И1. Кто сказал? Кто такое сказал/ без маски не пускать? И2. Роспотребнадзор. Что ещё? И1. Не знаю/ кто сказал. И2. Не знаю. Идите/ узнайте!» (конфликт в магазине, диалог 54); «И1. Дороговато содрали — семь тысяч. Ну/ оборудование/ счётчи ээ ящик. И2. Ага. И1. Тыща четыреста пятьдесят. Вот эти зажимы/ тридцать метров изолированного провода/ автоматы/ работа. Ну/ материалы — три тыщи пятьсот пятьдесят. И2. Угу. И1. Счётчик я свой поставил» (разговор о плате за электричество, диалог 77) и др.

Количественные показатели, характеризующие длину реплик в естественном диалоге, в среднем составляют 4–5 слов (около 22–23 символов), что соответствует минимальным показателям длины реплик персонажа «Собеседник», который в сгенерированных диалогах выполняет роль пользователя, обращающегося к диалоговому агенту.

Наблюдение над характером переходов между репликами в естественных диалогах показывает следующее.

Лексические маркеры переходов выражены менее ярко, чем в сгенерированных текстах, в частности, это показывает, меньшую структурно-грамматическую связь реплик, образующих диалогическое единство. Например, в искусственных диалогах наиболее частотный маркер перехода — обмен приветственными репликами, в то же время для естественной коммуникации

реальный лес/ реальные люди/ реальные инопланетяне/ скажем так. Вот. Также мне очень нравится/ как там проработаны персонажи/ даже второстепенные» (интервью о любимом фильме, книге, игре и музыке, диалог 26); «Вячеслав А. Мне вот про Джеймса Бонда нравятся фильмы после того/ как Дэниел Крейг начал сниматься в роли Джеймса Бонда. Это вот как-то заново я первый раз увидел — франшизе-то на самом деле/ бондиана/ как её называют/ ей уже несколько десятков лет — я увидел первый раз в середине восьмидесятых годов в кинофильм этот/ назывался он "Никогда не говори никогда". Шон Коннери тогда играл Джеймса Бонда. Я сюжет-то шас уже почти не помню» (монолог о фильмах, диалог 31) и др.

обмен приветствиями не является эталонным типом перехода к следующей реплике в ситуации общения. Распределение переходов, содержащих стандартные маркеры приветствия (например, *привет, здравствуйте*) в начале беседы для естественных диалогов, приведено на рис. 29 и составляет не более 20% от общего числа примеров.

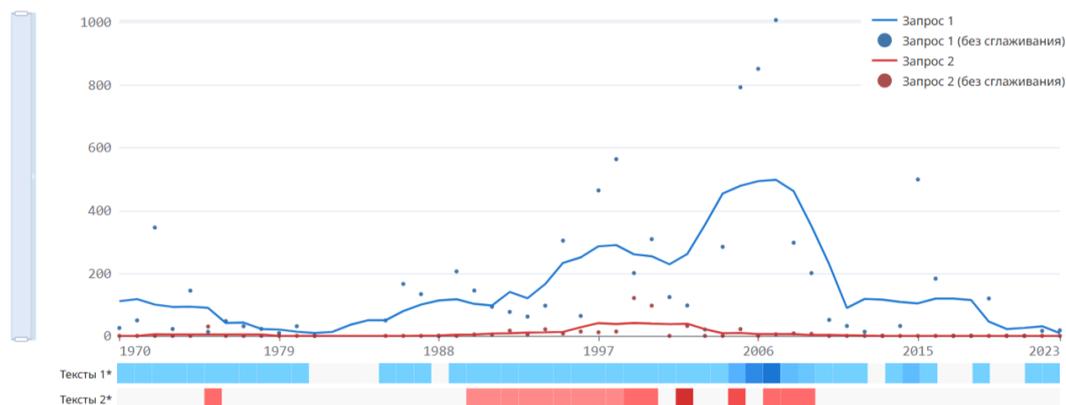


Рисунок 29. Частота вхождения реплик, содержащих эталонные фразы приветствия

Анализ частотных паттернов перехода между репликами в естественных диалоговых текстах позволяет выделить следующие их типы (по степени частотности):

ответы на вопросы собеседника утвердительного или отрицательного характера, выражение согласия или несогласия с мнением собеседника, эмоционально-оценочные реакции с семантикой подтверждения, согласия, одобрения (лексико-грамматические средства выражения: «да», «нет», «угу», «ага», «ну», «конечно», «точно», «вот и я», «неважно», междометий, вводных слов наподобие «значит», «думаю» и проч.);

комментарии к репликам собеседника, сообщения уточняющего характера (выражаются с помощью повтора частиц «да-да», вопросов «А где / как/ когда / почему ты...?», «А ты не...?», уточняющих вопросов к реплике: «До куда?», «До музыкалки?», «До станции?», контактоустанавливающих элементов наподобие «ты знаешь», «вот видишь» и др.);

сообщение новой информации или её запрос (используются различные типы речевых актов, преимущественно репрезентативы, изредка – комиссивы – обещания, рогативы – вопросы);

этикетные реплики (приветствие, прощание, благодарность), которые выражаются стандартно с использованием лексем «привет», «здравствуйте», «здорово», «пока», «до свидания», «спасибо».

Типология переходов между репликами в естественных диалогах носит более условный характер, чем в сгенерированных текстах, так как речевые реакции коммуникантов носят смешанный характер. К примеру, запросу или сообщению информации всегда соответствует речевая экспрессия, а также оценка содержания реплик собеседника, речевое выражение их одобрения или неодобрения. В связи с этим процентное соотношение не позволяет получить в сумме 100% из-за пересечений между типами речевых актов в репликах.

Количественное соотношение выявленных типов речевых реакций в естественных диалогах, которые маркируют переход между репликами, представлено на рис. 30.

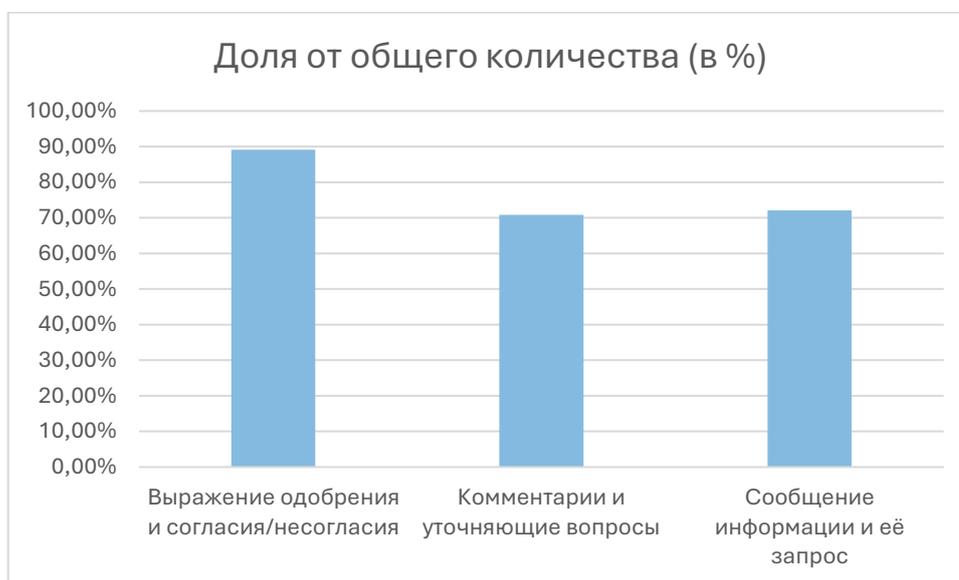


Рисунок 30. Количественное распределение переходов между репликами по типам речевых актов (естественный диалог)

Сопоставление с типами переходов между репликами, выявленными в сгенерированных диалоговых текстах, позволяет говорить о значимых различиях между структурой естественного и сгенерированного текста:

естественные диалоги характеризуются многообразием типов переходов между репликами, которые, несмотря на наличие рекуррентной связи реплик, демонстрируют синкретизм в выражении прагматических установок в речи;

сгенерированные диалоговые тексты следуют предопределённым структурам диалогового взаимодействия, ограничивающим вариативность реплик и приводящим к заданным паттернам развертывания коммуникации;

в естественных диалогах взаимодействие между спикерами носит более активный характер и характеризуется высокой степенью вовлеченности коммуникантов, что маркируется экспрессивами, неречевыми реакциями на реплику собеседника (улыбка, смех, вздох);

взаимодействие между спикерами в сгенерированных диалоговых текстах распределяется неравномерно, инициатива исходит от спикера, который выполняет роль пользователя алгоритма, а собеседник включается во взаимодействие в основном в пассивной роли – как информант, консультант, помощник, предоставляющий информацию по запросу в ходе беседы;

диапазон речевых средств, используемых для выражения речевой экспрессии в естественных диалогах, достаточно широк и не ограничивается эмоционально-оценочной лексикой, включая, например, изменение порядка слов или интонационного рисунка фразы;

в синтетических диалогах речевые маркеры эмоциональной реакции на реплику собеседника упрощены, однотипны, представлены повторяющимся из диалога в диалог набором лексико-грамматических средств (частицы, междометия);

элементы переходов между репликами в естественных диалогах контекстуально (и ситуативно) насыщены, поэтому реплики могут носить неразвернутый характер, но успешно выполнять коммуникативную функцию; с этим свойством также связано и отсутствие жесткой рекуррентной связи двух

соседних реплик: в целом, продолжая тему беседы, спикер может вводить новые микротемы, которые не всегда соответствуют содержанию предыдущей реплики собеседника¹⁹;

в искусственных диалоговых текстах, как правило, реплики, маркирующие смену спикеров, развернуты, приближаются по своему функционалу к монологу на тему и ориентированы не на контекст ситуации общения, а на содержание предыдущей реплики.

Таким образом, формально-структурная модель сгенерированного диалогового текста представлена паттерном жесткой рекуррентной связи между соседними репликами, которая не позволяет диалогу выходить за рамки заданного сценария, предопределённого контекстом промта. Жесткая формальная и содержательная связь реплик в синтетических диалоговых текстах приводит к обеднению их информационного содержания, сокращению обсуждаемых микротем, отсутствию глубины в их развертывании, минимальной связью с контекстом, которая проявляется в наличии примет времени, ситуативно обусловленных фраз, маркеров припоминания информации, характерных для естественных диалогов.

Лексико-грамматическая организация естественных текстов – параметр, который позволяет обнаружить разговорную природу естественных диалогов по сравнению с синтетическими диалогами.

Лексическая система естественных диалогов характеризуется преобладанием разговорной лексики. Составление конкордансов в подкорпусе

¹⁹ Примеры из естественных диалогов: «Наталья Владимировна В клуб. А так куда ты хочешь? И2 Я чё-то опять в Питер захотела. У меня уже как будто как традиция — ездить туда каждое лето» (диалог 16); «Ксения Олеговна А. Можно ходить в студии танцевальные. Александра У. Ну/ тут надо с комплексами бороться/ Ксюш. Мне лень этим заниматься» (диалог 23); «Дмитрий Т. Зато бы научились дороги строить за столько тысяч лет. Даниил Л. Кстати/ ты вот говоришь/ дороги. Я недавно видел/ точнее/ не знаю/ када я видел/ строились римские дороги/ выкапывались... выкапывалась траншея там метра два» (диалог 44); «А. Но/ я тоже помню что мы так ужинали/ картошка/.. В Картошку/ масло топили/ растапливали... у нас не было так/ что всё время мясо/ да?.. У нас чаще такой был/ ужин/ картошка с маслом [нрзб] А. Ты до какого числа говоришь в отпуске?..» (диалог 98) и др.

текстов для различных типов разговорных языковых единиц позволяет установить их частотность и высокую степень встречаемости для различных диалогов (*щас, халупа, алкаш, рваться* в значении «стремиться», *топать, жрать, бухать, валить* в значениях «уезжать», «убивать», «обвинять в чем-либо», *шмотки, мотаться*; слов-паразитов *типа, значит, короче, это самое, ну вот* и проч.). При этом ни одна из указанных единиц не встречается в сгенерированных текстах, которые по лексическому составу не являются разговорно-экспрессивными.

Количественные показатели анализа лексики (индекс лексического разнообразия) показывает, что в естественных диалогах он в среднем составляет 0,44. По данным исследований [Журавлев 1988, с. 16], для разговорных диалогических текстов в среднем показатель лексического разнообразия не превышает 0,54. Для сгенерированных диалогов он выше по отдельным темам, но в среднем также не превышает 0,58, что сопоставимо с лексическим богатством разговорного диалога. Следовательно, основные различия в лексическом составе между сгенерированными и естественными диалогами лежат не в количественном плане, а в качественном (состав лексики, её функционально-стилевая окраска, используемые тематические группы).

Синтаксис естественных диалогов может быть условно сопоставлен с синтаксической организацией сгенерированных диалоговых текстов ввиду разной функциональной направленности рассматриваемых текстов. Естественные диалоги, бытующие в устной непубличной сфере, характеризуются разговорным синтаксисом, включая:

нарушение прямого порядка речи, инверсию (например: «У нас у ДК Димка сказал лучше было гулянье», диалог 2; «И вот он там стажировался значит в Вене», диалог 9, «Ведь это же ужасно трудно / он будет тут мотаться», диалог 11 и др.);

дистантное расположение синтаксически связанных членов предложения, например, компонентов составных глагольных и именных сказуемых («Он по размеру/ наверно/ ну может/ чуть выше рассела он был», диалог 3; «Пьёт/ пьёт/ да

тут уже стал он меня бить и всё», диалог 42); членов субстантивных сочетаний («мне там Наташкины очень нравятся фотки», диалог 5);

препозиция ремы («Чтоб тебе йогурт купить питьевой на завтрак что-то забыла», диалог 49; «Тыща четыреста пятьдесят. Вот эти зажимы/ тридцать метров изолированного провода/ автоматы/ работа», диалог 77);

конструирование предложений по типу ассоциативного присоединения компонентов конструкции («И я шла/ шла/ шла/ я такая думаю/ ну хотя бы один бы упал бы/ абрикос/ и через секунду я вижу/ что что-то упало на пол/ и когда я увидела/ что это действительно абрикос/ я была просто в шоке/ что как это может произойти/ и я обрадовалась/ показала его всем», диалог 85; «А потом стала когда поливать/ жарко/ а ему ж приятно/ да ещё/ чешу/ гребнем/ стоит балдеет/ хорошо ему нравится/ водичкой поливаю/ а потом мылом или/ шампунем...», диалог 114).

Синтаксическая организация синтетических текстов тяготеет к структурным особенностям письменной речи, что выражается в уже выявленных нами синтаксических особенностях сгенерированных диалогов (преимущественно прямой порядок слов, отсутствие ассоциативного «нанизывания» придаточных предложений, вставных конструкций, отсутствие самоперебивов и уточнений, нарушающих нормы построения предложений в письменной речи).

В целом сопоставление синтаксической структуры естественного и искусственного диалогического текста является непрезентативным, поскольку большие языковые модели изначально обучаются на корпусе письменных текстов, поэтому воспроизводят только те паттерны устной коммуникации, которые имеют имитационную природу, то есть являются стилизацией под разговорную речь.

Обобщая сказанное, можно отметить, что лексико-грамматическая организация синтетических диалоговых текстов тяготеет к письменной монологической речи или литературному диалогу, имитирующему отдельные стилистические приемы, используемые в устной речи. Поскольку создание сгенерированных диалогов имеет конкретную прагматическую установку, связанную с развитием способности системы к поддержанию релевантного диалога с пользователем в определённой сфере, коммуникативной необходимости

в воспроизведении морфолого-синтаксических характеристик естественной диалоговой речи не возникает. Ситуации «фатического» взаимодействия пользователя и диалогового агента также нельзя признать аналогом естественной неформальной коммуникации, поскольку пользователь осведомлен о технологической природе агента и отсутствии у него релевантного коммуникативного опыта для полноценного общения. Поэтому такие диалоги можно рассматривать как спонтанные коммуникативные эксперименты, расширяющие опыт взаимодействия субъекта с современными сложными системами на основе искусственного интеллекта, что в целом соответствует концепции метаязыковой рефлексии и игрового освоения возможностей ИИ-агентов [Плотникова 2023].

Обобщая результаты проведенного исследования, представим результаты сравнительного анализа сгенерированных диалоговых текстов и естественных диалогов представлены в Таблице 16.

Таблица 16. Результаты сравнения искусственных и естественных диалоговых текстов

Компоненты анализа	Естественные диалоги	Сгенерированные диалоги
Тематика	Широкий спектр тем (включая общественно-политическую, экономическую, культурную) с доминированием тематики частной жизни	Тематический состав ограничен преимущественно кругом тем, касающихся знакомства, установления контакта и обсуждения увлечений, хобби (домашние животные, спорт)
Цель взаимодействия	Социальные интеракции, обмен опытом, выражение собственного мнения, оценка коммуникативной ситуации	Обучение взаимодействию с пользователем, выполнение задач текстогенерации, связанных с ответом на информационный запрос
Характер	Коммуникативная роль	Коммуникативная роль участника

Компоненты анализа	Естественные диалоги	Сгенерированные диалоги
коммуникативных ролей участников диалога	участника диалога определяется контекстом ситуации и формой социального взаимодействия, высокая степень вовлеченности участников в диалогическое взаимодействие, ситуативная насыщенность диалогов	диалога определяется заданным промтов паттерном, инициатива в диалоге в основном принадлежит пользователю (персонаж «Собеседник»), диалоговый агент («Ты») выполняет роль консультанта, помощника, содержание диалога не имеет ситуативной привязки
Структурные особенности	Отсутствие сценарного плана взаимодействия, ослабление рекуррентной связи между репликами в диалогическом единстве, короткие реплики	Наличие четкой структуры диалогового взаимодействия (приветствие – запрос информации – ответ), жесткая рекуррентная связь между репликами
Типы переходов между репликами	Многообразие типов переходов, смешанный характер речевых актов, синкретизм эмоционально-оценочных реакций, сообщения информации, уточнений	Стандартные маркеры переходов (взаимные приветствия, ответы на вопрос собеседника, выражение одобрения и оценки), ограниченная вариативность речевых реакций, их обособленность (одна реплика – одна речевая реакция)
Лексико-грамматическая организация диалогов	Преобладание разговорной лексики, использование слов-паразитов, нарушение порядка слов, паттерны разговорной синтаксической структуры	Преимущественно литературная лексика, отсутствие разговорных единиц, слов-паразитов, тяготение грамматической структуры к нормам к письменной речи
Информационная	Информационная избыточность	Минимальная глубина

Компоненты анализа	Естественные диалоги	Сгенерированные диалоги
насыщенность диалогов	сообщений, использование повторов, дублирование информации за счет уточнений, единство фактологической и оценочной информации	обсуждения темы, лаконичность содержания реплик, отсутствие повторов, уточняющих компонентов сообщения, преобладание фактологической информации над оценочно-экспрессивной
Контекстуализация	Связь с контекстом (эпохой, биографией коммуникантов, ситуацией общения)	Отсутствие в диалогах «примет времени», минимальные указания на ситуацию общения

Таким образом, сравнительный анализ сгенерированных и естественных диалоговых текстов позволяет выявить отличия в тематическом, структурно-композиционном и лексико-грамматическом аспектах. Выявленные отличия связаны с типовой формально-содержательным строением сгенерированного текста, которые предстает как рекуррентная модель текстопорождения в ситуации взаимодействия двух коммуникативных агентов – пользователя и нейросетевой маски. Заданный промптом контекст диалога придаёт сгенерированным текстам черты сценарного взаимодействия, осуществляющего по определённым правилам с целью решения поставленной задачи текстогенерации (например, предоставление информации или рекомендаций по теме запроса).

3.6. Методика атрибуции сгенерированного диалогового текста на основе лингво-информационной модели

Под атрибуцией искусственного диалогового текста мы понимаем процесс идентификации текста, созданного с помощью нейросетевых алгоритмов и диалоговых агентов.

Методика атрибуции сгенерированного диалогового текста предстает как инструментарий лингвистического исследования текста, позволяющий выявить паттерны текстогенерации с опорой на компоненты лингво-информационной модели синтетических диалогов, представленные нами в п.3.1-3.5 настоящего диссертационного исследования.

Следует отметить, что процесс генерации текста является сложным явлением, в котором возможны различные сценарии взаимодействия субъекта и диалогового агента, влияющие на качественные характеристики получаемого в результате генерации информационного сообщения. Так, при взаимодействии с диалоговыми системами часть реплик в диалоговом тексте принадлежит субъекту, а часть – алгоритму, что делает текст неоднородным в аспекте источника его происхождения, сочетающего два начала – антропологическое и технологическое. В случае полной генерации текста алгоритмом невозможно исключить из процесса текстопорождения процедуру промт-инжиниринга, поскольку промт является источником задачи текстовой генерации. Разная степень редактуры сгенерированного текста субъектом по окончании процесса генерации также вносит в результат работы алгоритмов существенные изменения, которые могут быть направлены на маскировку результатов генерации, придание тексту антропологических характеристик, изменение его структуры, содержания или речевого оформления. Указанные факторы обуславливают трудности, возникающие в процессы атрибуции сгенерированных текстов, и не позволяют решить эту проблему однозначно, даже с использованием программно-аппаратных средств. В связи с этим целесообразно говорить о выявлении лингвистических признаков генерации в исследуемом тексте, которые ставят под сомнение его авторство.

При разработке методики выявления признаков генерации в диалоговом тексте мы опираемся на лингвистические методы анализа, избегая применения методов машинного и (или) глубокого обучения, так как, на наш взгляд, ключевое различие между естественными и искусственными текстами диалогической природы лежит не в сфере количественных параметров (которые в целом

находятся в сопоставимых диапазонах), а в качественном своеобразии их содержания, структуры и речевого оформления.

Процедурные компоненты методики выявления признаков генерации в диалоге включают ряд этапов:

сбор данных для проведения атрибутивного анализа;

оценку пригодности объектов исследования для атрибуции;

комплексный тематический, семантический, лексико-грамматический и структурно-композиционный анализ исследуемых текстов для выявления их формальных и содержательных особенностей.

Для анализа необходимо выбрать репрезентативный корпус сгенерированных диалогов, поскольку единичный текст не является достаточным для проведения анализа в силу ограниченной возможности установления лингвистических паттернов сгенерированного сообщения.

Оценка пригодности исследуемого объекта для атрибуции связана с определением его количественных и качественных характеристик, релевантных для анализа. В частности, исследованию подвергаются количественные показатели объема текста, совокупности реплик, а также качественные параметры, определяющие тематическое единство, целостность текста, его формально-грамматическую и смысловую связность. Как показывают данные, полученные в ходе настоящего диссертационного исследования, критерии пригодности текста для атрибуции включают следующие параметры (представлены в Таблице 17).

Таблица 17. Критерии пригодности текста для проведения атрибутивного анализа на предмет признаков генерации

Параметр	Допустимое значение параметра
Объем реплик в диалоге	3 и более
Количество текстов в корпусе диалогов	5 и более
Тематическое единство диалога	Наличие единой темы, связывающей реплики в диалоге

Параметр	Допустимое значение параметра
Целостность текста	Текст диалога отличается связностью и завершенностью
Формально-грамматическая связность диалога	Текст диалога характеризуется показателями лексико-грамматической и формально-структурной связи между репликами
Смысловая связность диалога	Компоненты текста логически связаны, непротиворечивы, обусловлены развитием основной темы диалога

Темы диалогов могут быть выделены с помощью лексического анализа с применением контент-анализа для определения частотности тех или иных тематических паттернов.

Лексико-грамматический анализ диалогов включает:

подсчет частотности слов определенных частей речи;

расчет некоторых статистических метрик (средней длины реплик, индекса лексического разнообразия, индекса Флеша).

Анализ структурных особенностей диалоговых текстов предполагает:

исследование маркеров переходов между репликами (приветствие, запрос информации, сообщение информации, оценочные речевые реакции, уточняющие вопросы и проч.), что позволяет определить, как структурируется диалоговое взаимодействие;

создание типологии переходов между репликами для выявления рекуррентной модели диалогового общения.

Результаты тематического, структурно-композиционного и лексико-грамматического анализа сопоставляются для выявления закономерностей в репрезентации информационного содержания и атрибуции текста как обладающего признаками сгенерированного сообщения.

Предложенная методика может быть апробирована для анализа диалогов, сгенерированных специализированной нейросетью для создания диалоговых текстов Free AI Dialogue Generator (<https://www.writecream.com/free-ai-dialogue-generator/>). Для апробации методики с помощью указанной нейросети нами было сгенерировано 20 русскоязычных диалогов, отвечающих указанным в методике количественным параметрам пригодности, по запросу (промту) (приведен на английском языке): «10 dialogues in Russian, aiming for an informal, friendly tone between teenage readers discussing Harry Potter, sharing opinions and impressions» («10 диалогов на русском языке в неформальном дружеском тоне между читателями-подростками, которые обсуждают книги и фильмы, обмениваются мнениями и впечатлениями»). Примеры диалогов приведены в Приложении 4.

Тексты диалогов объединяются единой темой, связанной с обсуждением персонажей книг и фильмов о Гарри Поттере (серия книг «Гарри Поттер» английской писательницы Дж. К. Роулинг). Все диалоги характеризуются связностью и завершенностью. Темы развиваются последовательно, фактологические ошибки, противоречивые фрагменты (реплики) в диалогах отсутствуют. Реплики (реплика 1 – реплика 2 – реплика 3) логически связаны и образуют целостное в смысловом и формальном отношении диалогическое единство. Реплики структурированы, не содержат явных грамматических ошибок, аграмматизмов, речевых оборотов, не характерных для русской речи.

Тематический анализ диалогов показывает, что лексико-тематическое поле корпуса текстов ограничено тематическим паттерном, заданным в запросе на генерацию, отступлений от темы нет. Лексико-семантические показатели введения новых микротем отсутствуют, речевые маркеры, указывающие на личный опыт спикеров, связанный с обсуждаемой темой, а также маркеры ситуативной «привязки» диалогов не выявлены (обстоятельства, связанные с чтением книг, просмотром фильмов и проч. в диалогах не упоминаются).

Результаты формального (количественного) анализа диалогов представлены в Таблице 18.

Таблица 18. Средние значения количественных показателей диалогов (в разрезе реплик спикеров)

Тип спикера	Средняя длина предложения	Индекс Флеша	Количество вопросительных предложений	Количество восклицательных предложений	Количество повествовательных предложений
Спикер, открывающий диалог	5,94	116,2	0,43	1,4	3,5
Спикер-собеседник	5,52	116,6	0,14	0,7	2,9

Приведенные количественные данные показывают четкое распределение определённых формальных паттернов диалогового взаимодействия в зависимости от роли спикера в диалоге. Так, спикер, инициирующий диалог, как правило, является источником информации (доля повествовательных предложений в его репликах в среднем превышает долю аналогичных реплик спикера-собеседника). В репликах спикера-инициатора диалога также больше восклицательных и вопросительных предложений, поскольку он не только сообщает какую-либо информацию, задающую тематику беседы, но и маркирует свое к ней отношение оценочными репликами или вопросами. При этом второму спикеру в диалоге отводится пассивная роль (воспринимает сообщаемую информацию, выражая свое согласие или несогласие, отвечает на вопросы собеседника, уточняет и дополняет ответы). В целом формальные показатели структуры диалогов соответствуют выявленным в п. 3.3 настоящего исследования количественным закономерностям, характеризующим взаимодействие спикеров в сгенерированных диалоговых текстах, где одному из коммуникантов принадлежит инициатива, а второй играет пассивную роль с речевыми маркерами реактивного типа.

Анализ частеречной принадлежности лексики диалогов, а также синтаксических маркеров позволяет установить, что в текстах преобладают имена, включая существительные (в основном имена собственные) и прилагательные, также высока доля знаков препинания. Это соответствует паттернам сгенерированного диалога, ориентированного на обсуждение предметной действительности, фактов, идей, в то время как в естественных диалоговых текстах доля существительных и глаголов примерно одинакова, что свидетельствует об обсуждении событий, действий и состояний. Доля местоимений в исследуемых текстах также незначительна, что в целом характерно для сгенерированных диалогов, но не для естественной диалоговой коммуникации, где индекс дейктичности достаточно высок в силу ситуативности таких взаимодействий (рис. 31).

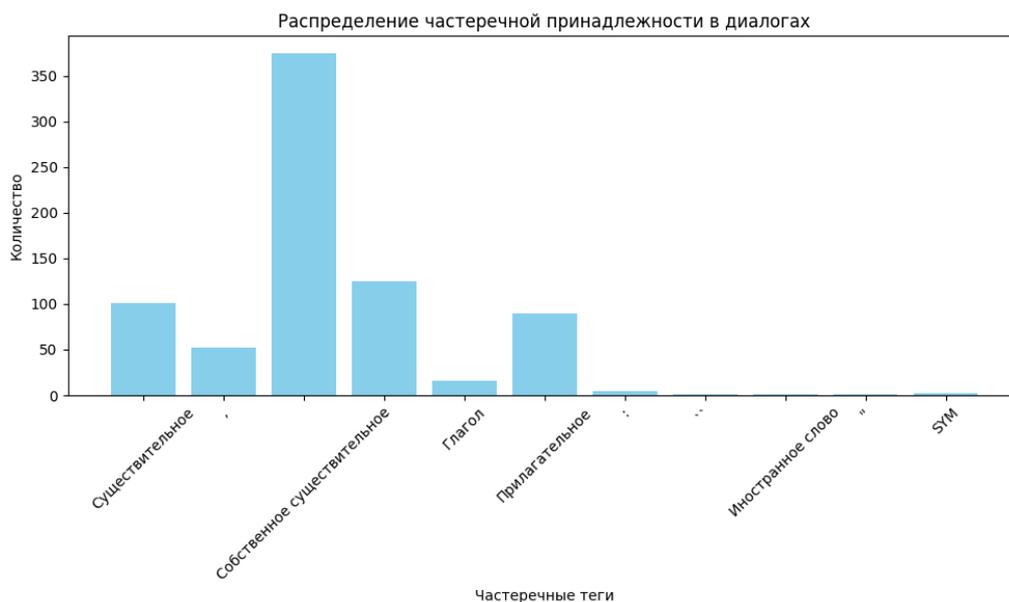


Рисунок 31. Распределение слов разных частей речи в диалогах

Результаты анализ структурных особенностей исследуемых текстов по типам переходов между репликами в диалогах приведены в Таблице 19.

При обращении к исследованию типов переходов между репликами выявлено, что речевые реакции собеседников в диалогах характеризуются высокой степенью предсказуемости. Так, наиболее частотным из наблюдаемых типов ответных реплик является переход от категории «информационное сообщение» к речевой реакции «комментарий» (46 переходов). Данный тип

перехода свидетельствует о существовании устойчивого рекуррентного паттерна в структуре диалога: получение информации инициирует реакцию собеседника в виде комментария к предшествующему сообщению.

Таблица 19. Переходы между категориями реплик (количественный анализ)

Предшествующая реплика	Последующая реплика	Количество
ответ	комментарий	1
комментарий	ответ	1
комментарий	информационные сообщения	2
комментарий	комментарий	2
информационные сообщения	ответ	9
информационные сообщения	запрос информации	11
информационные сообщения	информационные сообщения	19
информационные сообщения	комментарий	46

Частотные переходы внутри категории «информационные сообщения» (19) указывают на возможность повторного обращения к уже обсуждаемым темам для уточнения, что маркирует наличие в структуре исследуемых диалоговых текстов рекуррентного паттерна диалоговой модели: инициация общения – запрос информации – ответ – уточнение информации.

Наоборот, единичные переходы между категориями реплик «ответ» к «комментарий» (1) и от категории реплик «комментарий» к типу «ответ» (1) демонстрируют, что для диалогов практически не характерно введение новой информации в ответ на замечание (комментарий) собеседника или инициация углубления темы в ходе обмена репликами. Такой диалог лишен внутренней динамики, возникающей на основе активного участия обоих спикеров в развертывании темы обсуждения.

Типы ответных речевых реакций собеседников однотипны, повторяются из диалога в диалог, сводятся к одобрению реплик собеседника или уточнению информационного содержания предыдущей реплики. Экспрессивная и функционально-стилевая окраска реплик сглажена, что не позволяет установить особенности идиостиля собеседников. Однообразие средств речевой экспрессии для выражения оценочных суждений (преимущественно общеупотребительные прилагательные и существительные с оценочной семантикой: *умная, классный, крутой, гений* и др.) свидетельствует о низкой эмоциональной вовлеченности собеседников в тему обсуждения.

Выявленные особенности лингво-информационной модели исследуемых текстов включают:

тематическую ограниченность диалогов, отсутствие маркеров углубления темы и (или) отклонения от темы в ходе обсуждения;

контекстуальную автономность содержания диалогов от ситуации общения и опыта участников диалогического взаимодействия;

жесткую рекуррентную структуры диалогового взаимодействия, которое строится по модели «запрос / сообщение информации (стимул) – ответ на запрос / комментарий на основе предыдущей реплики (реакция)»;

устойчивые паттерны лексико-грамматической и речевой организации, не содержащие элементов разговорного стиля, маркеров индивидуальной речевой манеры собеседников;

избыточную правильность речи, литературность используемых конструкций, отсутствие орфографических ошибок, опечаток, неверной расстановки знаков препинания, которые характерны, например, для естественной письменной речи.

Выявленные особенности в соответствии с предложенной методикой атрибуции сгенерированного диалогического текста позволяют говорить о наличии признаков генерации в исследуемых диалоговых сообщениях.

Результаты оценки сгенерированных диалогов сервисами детекции сгенерированного текста представлены в Таблице 20.

Таблица 20. Результаты автоматической атрибуции исследуемых диалогов

Наименование сервиса	Результат проверки
Система «Антиплагиат» (https://users.antiplagiat.ru/)	Признаки AI-генерации отсутствуют
Neural Writer (https://neuralwriter.com/ru/content-detector-tool/)	Текст написан человеком с вероятностью 95%
Dupli Checker (https://www.duplichecker.com/ru/ai-content-detector.php)	Текст написан человеком с вероятностью 89,6%
AI Detector Writer (https://aidetectorwriter.com/ru/)	Текст написан человеком с вероятностью 100%
Decopy.ai (https://decopy.ai/ru/ai-detector/)	Текст создан человеком. Вероятность, что текст сгенерирован нейросетью, составляет 24%
YesChat (https://www.yeschat.ai/ru/features/ai-detector)	Текст с высокой вероятностью написан человеком
Smodin (https://smodin.io/ru/ai-детектор-контента)	Текст написан человеком с вероятностью 100%
Undetectable AI (https://undetectable.ai/ru)	Текст написан человеком с вероятностью 99%
Ahrefs (https://ahrefs.com/ru/writing-tools/ai-content-detector)	Текст написан нейросетью с вероятностью 80%
Sapling (https://sapling.ai/ai-content-detector)	Вероятность сгенерированного контента составляет 0%

При сопоставлении результатов автоматической проверки диалогов большинством доступных сервисов для детекции сгенерированного текста и результатов, полученных в ходе применения предложенной методики атрибуции,

можно сделать вывод о ненадёжности существующих автоматических сервисов обнаружения признаков текстовой генерации в диалогах и необходимости обращения к ручному лингвистическому анализу на основе комплексного подхода к структурно-содержательным характеристикам исследуемых текстов.

Выводы по третьей главе

Актуальность взаимодействия антропологического субъекта и систем на базе генерационных алгоритмов приводит к развитию коммуникации с диалоговыми агентами как особой формы обмена информацией между человеком и машиной. Успешность взаимодействия субъекта и диалогового агента, использующего алгоритмы генерации сообщений на естественном языке, требует учета лингвопрагматических аспектов, включая анализ специфики речевых актов и механизмов, способствующих эффективной коммуникации, которая определяется не только смысловой полнотой создаваемых сообщений, но и субъективной удовлетворенностью пользователей результатами работы систем генерации.

Лингво-информационное моделирование больших массивов сгенерированных диалоговых текстов как продуктов коммуникативного взаимодействия диалогового агента и субъекта охватывает аспекты тематического моделирования, структурно-композиционного и морфолого-синтаксического анализа, которые могут быть автоматизированы с использованием языка программирования Python. Полученные в ходе автоматического исследования формальные данные, подкрепленные качественным ручным анализом, демонстрируют ряд устойчивых признаков (паттернов) сгенерированного диалогового текста, к которым можно отнести:

ограниченное тематическое разнообразие, отсутствие глубины развертывания темы и микротем диалогов, формальную или отсутствующую связь содержания диалогов с контекстом, ситуацией общения и персональным опытом спикера;

низкий процент маркеров субъективной модальности в коммуникации, включая средства выражения речевой экспрессии, апелляции к коммуникативным намерениям и волеизъявлению спикера;

сценарную структуру диалогов, воспроизводящую ограниченный набор коммуникативных актов реактивного типа (приветствие – запрос информации – предоставление информации по запросу – уточнение информации – завершение взаимодействия);

линейность структуры диалогов, отсутствие непредсказуемых речевых реакций на реплики собеседника, регулярную воспроизводимость однотипных переходов между репликами в диалогах разных тематических групп;

высокую частотность субстантивной лексики, связанную с фактологическим и описательным характером диалогов, низкий индекс дейктивности как следствие внеситуативного характера их содержания;

избыточность пунктуации, высокий процент корневых зависимостей (ROOT), которые указывают на синтаксическую независимость предложений сгенерированных диалогов, характерную для монологической речи;

отсутствие системности в возникновении языковых ошибок, которые отражали бы особенности идиостиля спикеров.

Сравнение сгенерированных диалогов с естественными диалогами выявляет значительные различия в тематике, структуре, лексико-грамматической организации и информационной насыщенности, подчеркивая шаблонность и предсказуемость сгенерированных диалоговых сообщений.

Разработанная на основе результатов проведенного исследования методика атрибуции сгенерированного диалогового текста подчеркивает важность качественного анализа, позволяя идентифицировать лингвистические признаки автоматической генерации текста. Это, в свою очередь, открывает перспективы прикладных исследований в сфере улучшения качества текстогенерации, а также способствует формированию теоретических подходов к изучению искусственного текста в сравнительно-сопоставительном аспекте.

ЗАКЛЮЧЕНИЕ

В процессе исследования были решены следующие задачи: определены лингвистические основы генерационных процессов текстопроизводства, лежащие в основе деривации текстов и механизмов трансформации языковых структур по формальным правилам преобразования, позволяющих моделировать процессы создания новых текстов на основе существующих языковых единиц и правил их комбинирования; определен феномен сгенерированного текста как объект лингвистического исследования, выявлены его отличия от цифрового текста и специфика методов его лингвистического анализа; выявлены и проанализированы стратегии текстопорождения в коммуникации с диалоговыми системами, включая аспекты промт-инжиниринга и оценки коммуникативного качества сгенерированного текста; исследованы принципы, лежащие в основе информационной модели содержания сообщения в контексте генерации текста, определяющие особенности его восприятия и понимания; проанализированы сгенерированные диалоговые тексты для выявления их характеристик и разработки лингво-информационной модели текста как продукта взаимодействия человека и машины; сопоставлены паттерны модели с примерами естественной диалогической коммуникации для выявления специфики их природы; предложена методика атрибуции сгенерированных диалоговых текстов на основе выявленных параметров без использования автоматических методов атрибутирования.

Результатом стало достижение цели исследования, заключающейся в выявлении и анализе лингвистических характеристик (параметров) сгенерированного текста, полученного в результате взаимодействия пользователя с системами языковой генерации на основе искусственного интеллекта, для разработки обобщенной лингво-информационной модели его содержания, отражающей особенности и закономерности данного типа текстов.

Лингво-информационное моделирование сгенерированных диалоговых текстов позволяет выявить устойчивые паттерны, отражающие их коммуникативно-лингвистическую природу, включая тематические паттерны,

структурную и морфолого-синтаксическую организацию. Проанализированные с помощью методов автоматической обработки естественного языка и квалификативных методов анализа, сгенерированные тексты были исследованы как комплексный феномен текстогенерации, основанный на стратегиях промт-инжиниринга. Единство в коммуникативной структуре сгенерированного текста промта (речевого стимула) и реакции в виде продукта генерации (ответа нейросетевого алгоритма) обуславливает устойчивую рекуррентную природу синтетического диалога, который отличается набором структурно-содержательных паттернов. Среди них можно выделить ограниченное тематическое разнообразие диалогов, отсутствие глубины раскрытия темы и связи с контекстом, недостаток речевой экспрессии, сценарный характер, линейность и предсказуемость переходов между репликами, избыточность пунктуации, изолированность и синтаксическую независимость реплик, непоследовательность в языковых ошибках, не выраженность признаков идиостиля.

Сравнение с естественными диалогами выявляет значительные различия в тематике, структуре и информационной насыщенности, подчеркивая шаблонность сгенерированных текстов. Методика атрибуции, основанная на исследовании, позволяет идентифицировать признаки автоматической генерации текста, открывая пути теоретического изучения искусственного текста и прикладного использования выявленных паттернов текстогенерации в атрибутивном анализе.

Полученные результаты исследования позволяют вынести на защиту следующие положения:

1. Генерация текста с использованием нейросетей интегрирует принципы трансформации и динамического преобразования текстовых структур, заложенные в теоретические модели деривации текста, на качественно новом техническом уровне.

Исторические корни теории деривации и порождающей грамматики, заложившие основу для понимания процессов текстообразования, восходят к идеям синергетической природы речепорождения В. фон Гумбольдта, А. А. Потебни. В генеративизме, который стал реакцией на экспансию системно-

структурного анализа текста, текст рассматривается как результат трансформации глубинных структур в поверхностные по правилам, заданным порождающей грамматикой. Постгенеративизм делает акцент на динамической (процессуальной) концепции текстопорождения. В теории деривации это связано с изучением изменений функций языковых элементов в процессе создания текста – ключевой идеей, положенной в основу современных подходов к обработке естественного языка, включая моделирования сложных текстовых структур. Современные нейросети, основанные на архитектуре трансформеров, эффективно реализуют принципы лексической и синтаксической деривации, обеспечивая многослойную обработку текстовых данных и учет структурно-семантических связей при развертывании текстового содержания. Автоматизированные системы, синтезируя подходы генеративной грамматики Н. Хомского, концепции «Смысл – Текст» И. А. Мельчука и идеи дериватологической теории текстообразования, на качественном новом техническом уровне позволяют добиться процессов текстопроизводства, соответствующих паттернам естественной коммуникации.

2. Сгенерированный текст представляет собой лингвистический объект нового типа, в котором антропологические стратегии текстопорождения заменяются алгоритмическими, что, наряду с автоматизацией создания и размножением текстов типовой структуры, параметризацией и прогнозированием содержания, приводит к отчуждению текста от речевой деятельности субъекта и утрате индивидуально-авторских черт.

Сгенерированный текст создается с помощью алгоритмов и нейросетевых генераторов на основе закономерностей естественного языка. Он отличается от цифрового текста, который может включать как сгенерированные, так и традиционные тексты, переведенные в цифровой формат, тем, что создается автоматически, с опорой на механизмы глубокого обучения и оптимизации текстовых структур.

Лингвистическая специфика сгенерированного текста заключается в гетерогенной семиотической природе, основанной на комбинации различных знаковых систем и их взаимодействии; автоматизации процессов создания;

отсутствием связи с контекстом речевой ситуации и механизмами семиотической трансформации индивидуального когнитивного опыта. Методология анализа сгенерированного текста, сочетая количественные и качественные методы, позволяет исследовать его как лингвистический объект, сопоставляя по ряду характеристик с естественными текстами для выявления специфических формально-содержательных признаков, отраженных в структуре и лексико-тематической организации. Оценка коммуникативного качества сгенерированного текста включает два аспекта – объективный и субъективный. Первый позволяет определять степень грамматической и семантической приемлемости текста, его соответствие нормам языка и коммуникативным ожиданиям, второй – релевантность текста целям и ожиданиям пользователя, а также его способность выполнять заданные функции в рамках процессов текстогенерации. При этом коммуникативные качества сгенерированного текста в целом соответствуют традиционным представлениям о коммуникативной эффективности речевого произведения и могут быть интерпретированы в категориях точности, ясности, логичности, выразительности.

3. Промт-инжиниринг как методология разработки и оптимизации текстовых запросов для взаимодействия с нейросетевыми алгоритмами улучшает коммуникативное качество продуктов текстогенерации и трансформирует механизмы организации информации и восприятия сгенерированных текстовых сообщений. Когнитивная модель интерпретации сгенерированного текста задается выбранной стратегией промт-инжиниринга в соответствии с его жанровым аспектом и концептуализируется с помощью пропозициональной и семантической структуры генерируемого сообщения. Запрос на генерацию текста (промт) выступает в качестве новой стратегии текстопорождения, возникающей в эпоху больших языковых моделей. Промт позволяет взаимодействовать с нейросетевыми алгоритмами на естественном языке без использования специальных навыков программирования, «гуманизируя» процесс взаимодействия субъекта с машинным алгоритмом. Промт, в отличие от команд структурированных языков запросов, обладает свободной структурой, что делает

результаты генерации менее предсказуемыми и в целом соответствующими моделям стохастического порождения речи. Промты могут быть классифицированы по различным критериям, включая тип задачи, уровень детализации и целевую аудиторию. Промт-инжиниринг представляет собой методику разработки и оптимизации текстовых запросов для улучшения взаимодействия между пользователем и нейросетевым алгоритмом. Включая техники тестирования различных вариантов промтов и использование шаблонов для повышения предсказуемости результатов, промт-инжиниринг позволяет прогнозировать формально-содержательные признаки генерируемого текста. Возможности формализации промта ограничены коммуникативными параметрами генерируемого текста и накладывают ограничения на степень его детализации и параметризации, требуя от пользователя реализации метаописательных стратегий при создании промтов.

Стратегии промт-инжиниринга включают когнитивную оптимизацию, позволяющую пользователям активно формировать промты для более эффективной генерации текста; интерактивность, обеспечивающую динамичное взаимодействие с алгоритмами; мультимодальность, расширяющую форматы представления генерируемой информации; технологии распределённого субъекта познания и коммуникации, делегирующие часть познавательных и интерпретационных полномочий алгоритмам.

Единство промта и порождаемого текста формирует диалоговую форму коммуникации, придавая продуктам текстогенерации интерактивность, динамизм, комплементарность и симультанность, становящиеся базовыми компонентами информационной модели содержания сообщения в контексте генерации текста. При этом меняется статус коммуникативных агентов: на смену антропологическим субъектам в коммуникации приходят распределённые эпистемические системы, формирующие лингвокогнитивные механизмы восприятия и интерпретации сгенерированных сообщений на основе лингвистических маркеров «гуманизации» искусственного текста.

4. Комплексная лингво-информационная модель сгенерированного текста включает тематические, структурно-композиционные и морфолого-синтаксические компоненты, позволяющие выявить специфические характеристики автоматического текстопроизводства, включая паттерны речевых актов, маркеры переходов между репликами и лексико-грамматическую организацию диалога. Сравнение сгенерированных и естественных диалогов по параметрам тематического разнообразия, частотности субстантивной лексики, значения индекса дейктивности, наличия маркеров речевой экспрессии, уровня формально-грамматической структурированности подчеркивает качественные различия между двумя типами текстов, выявляя

Лингво-информационная модель представляет собой теоретическую конструкцию, которая описывает, как текстовая информация организуется и передается через различные уровни структуры текста. Данная модель сгенерированного текста включает в себя несколько ключевых компонентов: формальный, лексический, семантический, прагматический. В ходе исследований были выявлены устойчивые характеристики, которые помогают отличить сгенерированные тексты от естественных. К ним относятся структурная однородность, рекуррентная природа, ограниченная тематическая вариативность, преобладание субстантивно-атрибутивных форм при низком уровне дейктивности и при минимальном числе грамматических маркеров речевой экспрессии, синтаксическая автономность диалоговых реплик, высокий уровень грамматической структурированности, отсутствие аграмматизмов, системных речевых сбоев, оговорок, ситуаций, моделирующих коммуникативные неудачи и проч. Данные паттерны, отражая устойчивые признаки генерации текстового содержания, могут быть положены в основу исследования базовых принципов моделирования информационно-содержательных и структурных сгенерированных текстов диалоговой природы. Сравнительный анализ сгенерированных и естественных диалогов показывает, что последние обладают большей глубиной, разнообразием и контекстуальной насыщенностью, что делает их более

эффективными для коммуникации. Это подчеркивает важность контекста и фоновых знаний в процессе взаимодействия.

5. Методика атрибуции сгенерированного диалогового текста основана на комплексном анализе, позволяющем выявлять признаки генерации. Процесс атрибуции включает оценку количественных и качественных характеристик текстов и позволяет получать более надежные результаты при атрибуции сгенерированного контента и (или) наличия его признаков тексте, чем при использовании автоматической проверки существующими сервисами обнаружения признаков автоматически созданного текста.

Атрибуция сгенерированного текстового контента предстает как процесс определения источника текста и его классификации как сгенерированного или естественного. Методика атрибуции носит поэтапный характер, включая сбор образцов сгенерированных и естественных текстов, выявление их ключевых параметров (лексико-грамматических, тематических, структурно-композиционных), сравнение на основе собранных данных для обнаружения различий между сгенерированными и естественными текстами на основе устойчивых паттернов лингво-информационной модели сгенерированного текста.

Результаты диссертационного исследования имеют теоретико-прикладное значение, позволяя использовать их для формирования методики теоретического анализа искусственных текстов, а также для атрибуции сгенерированного текстового контента или при разработке рекомендаций для создания более совершенных систем генерации текста. Нам представляется, что перспективы дальнейших исследований в данном направлении связаны с углублённым анализом различных типов и жанров сгенерированных текстов, механизмов их линвокогнитивного восприятия и аспекта влияния на коммуникацию в цифровом пространстве в целом. Значимым представляется также экспериментальный аспект текстогенерации, связанный с исследованием возможностей моделирования цифрового двойника личности для воссоздания его речевой манеры и решения проблемы «переноса стиля» в условиях автоматического копирования идиостилей в цифровом пространстве.

СПИСОК ЛИТЕРАТУРЫ

1. Автоматическая обработка текстов на естественном языке и анализ данных: учеб. пособие / Большакова Е. И., Воронцов К. В., Ефремова Н. Э., Клышинский Э. С., Лукашевич Н. В., Сапин А. С. – М.: Изд-во НИУ ВШЭ, 2017. – 269 с.
2. Агапова С. Г., Полоян А. В. Интернет-дискурс: основные жанры и особенности их исследования / С. Г. Агапова, А. В. Полоян // Известия Южного федерального университета. Филологические науки. – 2016. – №4. – С. 52–58.
3. Адливанкин С. Ю., Мурзин Л. Н. О предмете и задачах дериватологии / С. Ю. Адливанкин, Л. Н. Мурзин // Деривация и текст: межвуз. сб. науч. тр. / Перм. гос. ун-т. – Пермь, 1984. – С. 3–12.
4. Айдагулова А. Р. Особенности текстов, сгенерированных искусственным интеллектом / А. Р. Айдагулова // Вестник Башкирского государственного педагогического университета им. М. Акмуллы. – 2023. – №4 (72). – С. 154–156.
5. Алгоритмы предварительной обработки текста: декомпозиция, аннотирование, морфологический анализ / В. А. Яцко [и др.] // Научно-техническая информация. Сер. 2, Информационные процессы и системы. – 2009. №11. – С. 24–30.
6. Анисимова, Е. Е. Лингвистика текста и межкультурная коммуникация (на материале креолизованных текстов): учебное пособие для студентов факультетов иностранных языковых вузов / Е. Е. Анисимова. – Москва: Издательский центр «Академия», 2003. – 128 с.
7. Апресян Ю. Д. Интегральное описание языка и системная лексикография / Ю. Д. Апресян. – М.: Языки русской культуры, 1995. – 768 с.
8. Арутюнова Н. Д. Типы языковых значений: Оценка. Событие. Факт / Н. Д. Арутюнова. – М.: Наука, 1988. – 338 с.
9. Байкова А. В. Мультиmodalный дискурс-анализ как один из методов коммуникативно-прагматического подхода к пониманию языка и текста / А. В.

Байкова // Вестник Шадринского государственного педагогического университета. – 2022. – №2 (54). – С. 230–234.

10. Балашова И. Ю., Волынская К. И., Макарычев П. П. Методы и средства генерации тестовых заданий из текстов на естественном языке / И. Ю. Балашова, К. И. Волынская, П. П. Макарычев // Модели, системы, сети в экономике, технике, природе и обществе. – 2016. – №1 (17). – С. 195–202.

11. Баркович А. А. Интернет-дискурс. Компьютерно-опосредованная коммуникация: учебное пособие по интернет-лингвистике / А. А. Баркович. – Москва: Флинта, Наука, 2016. – 285 с.

12. Бартенев О. В. Оценка эффективности методов токенизации текста / О. В. Бартенев // Вестник Московского энергетического института. – 2023. – №6. – С. 144–156.

13. Батраева И. А., Нарцев А. Д., Лезгян А. С. Использование анализа семантической близости слов при решении задачи определения жанровой принадлежности текстов методами глубокого обучения / И. А. Батраева, А. Д. Нарцев, А. С. Лезгян // Вестник Томского государственного университета. Управление, вычислительная техника и информатика. – 2020. – №50. – С. 14–22.

14. Бахтин М. М. Проблема речевых жанров / М. М. Бахтин // М. М. Бахтин. Литературно-критические статьи. – Москва: Худож. лит., 1986. – С. 428–472.

15. Белов С. Д., Зрелова Д. П., Зрелов П. В. и др. Обзор методов автоматической обработки текстов на естественном языке / С. Д. Белов, Д. П. Зрелова, П. В. Зрелов, В. В. Кореньков // Системный анализ в науке и образовании. – 2020. – №. 3. – С. 1–15.

16. Белогорская Д. В., Резанова З. И. Лингвистическая оценка автоматически сгенерированных рефератов новостных текстов / Д. В. белогорская, З. И. Резанова // Язык и культура. – 2023. – №61. – С. 15–28.

17. Белоедова А. В., Кожемякин Е. А. Мультимодальная коммуникация в фокусе исследовательской рефлексии: проблемно-ориентированный подход /

А. В. Белоедова, Е. А. Кожемякин // Критика и семиотика. – 2022. – №2. – С. 54–70.

18. Белозерова Н. Н. Мир реальный и мир виртуальный: две экологические системы?: монография / Н. Н. Белозерова. – Тюмень: Изд-во Тюменского гос. ун-та, 2010. – 251 с.

19. Белозерова Н. Н., Чуфистова Л. Е. Когнитивные модели дискурса / Н. Н. Белозерова, Л. Е. Чуфистова. – 2-е изд. – Тюмень: ТюмГУ, 2013. – 256 с.

20. Белоусов К. И. Синергетика текста: от структуры к форме: монография / К. И. Белоусов. – М.: Книжный дом «ЛИБРОКОМ», 2008. – 248 с.

21. Белоусов К. И. Теория и методология полиструктурного синтеза текста / К. И. Белоусов. – М.: Флинта: Наука, 2009. – 216 с.

22. Бенгфорт Б., Билбро Р., Охеда Т. Прикладной анализ текстовых данных на Python. Машинное обучение и создание приложений обработки естественного языка / Б. Бенгфорт, Р. Билбро, Т. Охеда. – СПб.: Питер, 2019. – 368 с.

23. Богданова-Бегларян Н. В., Блинова О. В., Мартыненко Г. Я. И др. Некоторые инвариантные характеристики русской разговорной речи: фонетика, морфология, синтаксис / Н. В. Богданова-Бегларян, О. В. Блинова, Г. Я. Мартыненко, Т. Ю. Шерстинова // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной международной конференции «Диалог» (Москва, 31 мая — 3 июня 2017 г.). Вып. 16 (23): В 2 т. Т. 2. – М.: Изд-во РГГУ, 2017. – С. 75–86.

24. Большакова Е. И., Клышинский Э. С., Ландэ Д. В. и др. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика: учебное пособие / Е. И. Большакова, Э. С. Клышинский, Д. В. Ландэ, А. А. Носков, О. В. Пескова, Е. В. Ягунова. – М.: Изд-во МИЭМ, 2011. – 272 с.

25. Бондаренко И. В. Влияние генеративной лингвистики Н. Хомского на мировое языкознание / И. В. Бондаренко // Вестник Балтийского федерального ун-та им. И. Канта. – 2011. – Вып. 2. – С. 141–149.

26. Бонч-Осмоловская, А. А. Предсказания, большие данные и новые измерители: о возможностях технологий компьютерной лингвистики в теоретических лингвистических исследованиях / А. А. Бонч-Осмоловская // Вопросы языкознания. – 2016. – №2. – С. 100–120.

27. Борисова И. Н. Русский разговорный диалог. Структура и динамика / И. Н. Борисова. – М.: Книжный дом «ЛИБРОКОМ». – 320 с.

28. Бочаров В. В., Митренина О. В., Добров А. В. Прикладная и компьютерная лингвистика: коллективная монография / В. В. Бочаров, О. В. Митренина, А. В. Добров / под редакцией И. С. Николаева, О. В. Митрениной, Т. М. Ландо. – Москва, 2017. – 315 с.

29. Бурнашев Р. Ф. Роль экспертных систем в лингвистических исследованиях / Р. Ф. Бурнашев // Science and Education. – 2023. – Т. 4. – №. 3. – С. 941–950.

30. Бурнашев Р. Ф. Информационные технологии в решении проблем современной лингвистики / Р. Ф. Бурнашев // Universum: филология и искусствоведение. – 2023. – №6 (108). – С. 23–27.

31. Бурнашев Р. Ф., Аламова А. С. Роль нейронных сетей в лингвистических исследованиях / Р. Ф. Бурнашев, А. С. Аламова // Science and Education. – 2023. – Т. 4. – №3. – С. 258–269.

32. Бурнашев Р. Ф., Ахадова Ш. С., Нематуллаева Н. Б. К вопросу об особенностях лингвистических корпусов второго и третьего поколений / Р. Ф. Бурнашев, Ш. С. Ахадова, Н. Б. Нематуллаева // Европа, наука и мы: сборник научных публикаций международной научно-практической конференции. – Издательство «Education and Science». Чехия: Прага. – 2021. – С. 77–79.

33. Бурнашев Р. Ф., Мурзамуратова У. Б. Применение технологий компьютерной лингвистики в социальных сетях и интернет-маркетинге / Р. Ф. Бурнашев, У. Б. Мурзамуратова // Universum: филология и искусствоведение. – 2023. – №10 (112). – С. 14–19.

34. Бурнашев Р. Ф., Нематуллаева Н. Б., Худоярова П. Н. Роль лингвистических корпусов в научных исследованиях / Р. Ф. бурнашев, Н. Б.

Нематуллаева, П. Н. Худоярова // Science and education: сборник научных публикаций международной научно-практической конференции. Турция, Анталия. – 2021. – С. 126–128.

35. Бхат С. Б. Философия семиотики визуальной коммуникации в контексте массовой культуры / С. Б. Бхат // Гуманитарные и политико-правовые исследования. – 2019. – №3. – С. 14–30.

36. Ван Дейк Т. А., Кинч В. Стратегии понимания связного текста // Новое в зарубежной лингвистике: Когнитивные аспекты языка / Т. А. Ван Дейк, В. Кинч. – М., 1988. Вып. XXIII. – С. 153–211.

37. Ван Дейк Т. А. Язык. Познание. Коммуникация / Т. А. Ван Дейк / Пер. с англ. Сост. В. В. Петрова. – М.: Прогресс, 1989. – 312 с.

38. Вашунина И. В., Егорова Л. А., Рябова М. Э. Поликодовый гипертекст в мультязычном дискурсе межкультурных коммуникаций / И. В. Вашунина, Л. А. Егорова, М. Э. Рябова // Преподаватель XXI век. – 2017. – №2–3. – С. 305–317.

39. Вержбицкая А., Речевые акты / А. Вержбицкая // Новое в зарубежной лингвистике, выпуск XVI: Лингвистическая прагматика. / Составление и вступительная статья Н. Д. Арутюновой и Е. В. Падучевой, общая редакция Е. В. Падучевой – М.: Прогресс, 1985. – С. 251–275.

40. Винокур Т. Г. Говорящий и слушающий. Варианты речевого поведения / Т. Г. Винокур. – Москва: Наука, 1993. – 172 с.

41. Володин А. Ю. Digital Humanities (цифровые гуманитарные науки): в поисках самоопределения / А. Ю. Володин // Вестник Пермского университета. Серия: История. – №3(26). – 2014. – С. 5–12.

42. Ворошилова, М. Б. Методология, методы и методика анализа креолизованного текста / М. Б. Ворошилова // Эволюция лингвистической экспертизы: методы и приемы / Э. В. Будаев, М. Б. Ворошилова, Н. Б. Руженцева. – Екатеринбург: Уральский государственный педагогический университет, 2017. – С. 125–199.

43. Гак В. Г. Высказывание и ситуация. Проблемы структурной лингвистики / В. Г. Гак. – М., 1972. – С. 349–372.

44. Галичкина Е. Н. Типология речевых жанров сетевой компьютерной коммуникации / Е. Н. Галичкина // Известия Волгоградского государственного педагогического университета. – 2019. – №2 (135). – С. 97–100.

45. Галло Я., Алефиренко Н. Ф. Лингвистика текста: традиции и перспективы / Я. Галло, Н. Ф. Алефиренко // Филологический класс. – 2020. – Т. 25, № 3. – С. 23–38.

46. Гальперин И. Р. Текст как объект лингвистического исследования \ И. Р. Гальперин. – М.: Наука, 2006. – 144 с.

47. Гладкий А. В., Мельчук И. А. Элементы математической лингвистики \ А. В. Гладкий, И. А. Мельчук. – М.: Наука, 1969. – 192 с.

48. Голев Н. Д. Динамический аспект лексической мотивации / Н. Д. Голев / под ред. О. И. Блиновой; Алт. гос. ун-т. – Томск: Издательство Томского университета, 1989. – 249 с.

49. Голев Н. Д. Деривация и мотивация как формы оязыковлённой детерминации: инвариант и варианты / Н. Д. Голев // Очерки по лингвистической детерминации и дериватологии русского языка. – Барнаул: Изд-во Алт. ун-та, 1998. – С. 65–84.

50. Голев Н. Д. Лексико-деривационное пространство русского текста: структура, семантика, прагматика / Н. Д. Голев // Фатическое поле языка: Межвуз. сб. научных трудов. – Пермь: Перм. ун-т, 1998. – С. 24–33.

51. Голев Н. Д., Сайкова (Мельник) Н. В. Изложение, пародия, перевод ... К основаниям деривационной интерпретации вторичных текстов / Н. Д. Голев, Н. В. Сайкова (Мельник) // Языковое бытие человека и этноса: психолингвистический и когнитивный аспекты. Вып. 3. – Барнаул: Изд-во Алт. ун-та, 2001. – С. 20–27.

52. Гольдберг Й. Нейросетевые методы в обработке естественного языка / Й. Гольдберг / пер. с англ. А. А. Слинкина. – М.: ДМК Пресс, 2019. – 282 с.

53. Городецкий Б. Ю. Компьютерная лингвистика: моделирование языкового общения / Б. Ю. Городецкий // Новое в зарубежной лингвистике. Выпуск XXIV. Компьютерная лингвистика. – Москва: Прогресс, 1989. – С. 5–31.

54. Гречачин В. А. К вопросу о токенизации текста / В. А. Гречачин // Международный научно-исследовательский журнал. – 2016. – №6–4 (48). – С. 25–27.
55. Григорьева В. С. Когнитивно-прагматические аспекты конструирования дискурса / В. С. Григорьева // Вопросы когнитивной лингвистики. – 2007. – №2. – С. 46–57.
56. Григорьева И. В. Аспекты теории глубинных структур в работах Н. Хомского / И. В. Григорьева // Вестник ВГУ. Серия: Лингвистика и межкультурная коммуникация. – 2018. – №4. – С. 9–14.
57. Гудков В. Ю, Гудкова Е. Ф. N-граммы в лингвистике / В. Ю. Гудков, Е. Ф. Гудкова // Вестник Челябинского государственного университета. – 2011. – №24. – С. 69–71.
58. Дедова О. В. Лингвосомиотический анализ электронного гипертекстов: на материале русскоязычного Интернета: автореферат дис. ... доктора филологических наук: 10.02.01, 10.02.19 / Дедова Ольга Викторовна, 2006. – 48 с.
59. Диалогическая лингвистика: коллективная монография / под науч. ред. Н. Н. Шпильной; отв. ред. М. В. Воронец, Н. И. Тюкаева. – Барнаул: АлтГПУ, 2019. – 320 с.
60. Дискурс современных массмедиа в перспективе теории, социальной практики и образования: II Международная научно-практическая конференция. Актуальные проблемы современной медиалингвистики и медиакритики в России и за рубежом: II Международный научный семинар. Белгород, НИУ «БелГУ», 5–7 октября 2016 г.: сборник научных работ / под ред. Е. А. Кожемякина, А. В. Полонского. – Белгород: ИД «Белгород» НИУ «БелГУ», 2016. – 380 с.
61. Добросклонская, Т. Г. Медиалингвистика: теория, методы, направления / Т. Г. Добросклонская. – [б. м.]: [б. и.], 2020. – 180 с.
62. Ейгер Г. В., Юхт В. Л. К построению типологии текстов / Г. В. Ейгер, В. Л. Юхт // Лингвистика текста: материалы научной конференции в МГПИИЯ имени М. Тореца. – М., 1974. – С. 103–109.

63. Журавлев А. Ф. Опыт количественно-типологического исследования разновидностей устной речи / А. Ф. Журавлев // Разновидности городской устной речи: Сб. науч. тр. / Под ред. Д. Н. Шмелева, Е. А. Земской. – М., 1988. – С. 84–151.

64. Загидуллина М. В. Мультиmodalность: к вопросу о терминологической определенности / М. В. Загидуллина // Знак: проблемное поле медиаобразования. – 2019. – №1. – С. 188–188.

65. Загидуллина М. В. Современное состояние мультиmodalного анализа: к вопросу о перспективах метода / М. В. Загидуллина // Научный результат. Социальные и гуманитарные исследования. – 2023. – Т. 9. – № 1. – С. 84–99.

66. Захаров В. П. Поисковые системы Интернета как инструмент лингвистических исследований / В. П. Захаров // Русский язык в Интернете: сб. ст. – Казань: Отечество, 2003. – С. 48–59.

67. Захарова Е. Ю. Лексическое разнообразие текста и способы его измерения / Е. Ю. Захарова // Вестник Тюменского государственного университета. Гуманитарные исследования. Humanitates. – 2020. – Том 6. – № 1 (21). – С. 20–34.

68. Земская Е. А. Русская разговорная речь. Общие вопросы. Словообразование. Синтаксис / Е. А. Земская, М. В. Китайгородская, Е. Н. Ширяев. – Москва, 1981. – 276 с.

69. Земская Е. А. Русская разговорная речь: лингвистический анализ и проблемы обучения / Е. А. Земская. – Москва: Русский язык, 1979. – 239 с.

70. Земская Е. А. Политематичность как характерное свойство непринужденного диалога / Е. А. Земская // Разновидности городской устной речи. – Москва, 1988. – С. 234–240.

71. Зильберман Н. Н. Технологии виртуальных собеседников и формы речевого взаимодействия / Н. Н. Зильберман // Гуманитарная информатика. – 2009. – №5. – С. 80–85.

72. Зиновьева Е. С. Изучение дискурса в рамках медиалингвистики / Е. С. Зиновьева // Верхневолжский филологический вестник. – 2016. – №3. – С.73–80.
73. Зобнин А. И., Носырев Г. В. Морфологический анализатор MyStem 3 / А. И. Зобнин, Г. В. Носырев // Труды Института русского языка им. В. В. Виноградова. – 2015. – Т. 6. – С. 300–310.
74. Иорданская Л. Н. Автоматический синтаксический анализ / Л. Н. Иорданская; под общ. ред. А. А. Ляпунова, О. С. Кулагиной; Академия наук СССР, Институт математики Сибирского отделения [и др.]. – Новосибирск: Наука. Сибирское Отделение, 1967. – Том 2. Межсегментный синтаксический анализ. – 231 с.
75. Иорданская Л. Н. Коммуникативная структура и ее использование в системе текстовой генерации / Л. Н. Иорданская // Международный форум по информации и документации. 1992. – Т. 17, №2. –С. 15–28.
76. Иорданская Л. Н. Смысл и сочетаемость в словаре / Л. Н. Иорданская, И. А. Мельчук; ред. Е. Н. Саввина. – Москва: Языки славянской культуры (ЯСК), 2007. – 672 с.
77. Исхакова А. О. Метод и программное средство определения искусственно созданных текстов: автореферат дис. ... кандидата технических наук: 05.13.17 Исхакова Анастасия Олеговна, 2016. – 22 с.
78. Карасик В. И. Жанры сетевого дискурса / В. И. Карасик // Жанры речи. – 2019. – №1 (21). – С. 49–55.
79. Карасик В. И. Речевая коммуникация: дискурсивный аспект / В. И. Карасик // Грани познания. – 2013. – №1(21). – С. 23–33.
80. Карасик В. И. Языковой круг: личность, концепты, дискурс / В. И. Карасик. – Волгоград: Перемена, 2002. – 477 с.
81. Квантитативная лингвистика: сборник научных статей / Министерство образования и науки РФ, Смоленский государственный университет; [отв. ред. С. Н. Андреев]. Т. 4. – Смоленск: Изд-во СмолГУ, 2014. – 131 с.

82. Квантитативная лингвистика: сборник научных статей / Министерство образования и науки РФ, Смоленский государственный университет; [отв. ред. С. Н. Андреев]. Т. 5. – Смоленск: Изд-во СмолГУ, 2015. – 113 с.

83. Кибрик А. А. К вопросу о структуре мультимедийного дискурса / А. А. Кибрик // Образы языка и зигзаги дискурса: сборник научных статей к 70-летию В. З. Демьянкова / В. В. Фещенко (ред.). – М.: Культурная революция, 2018. С. 180–191.

84. Кибрик А. А. Мультимодальная лингвистика / А. А. Кибрик // Когнитивные исследования: сборник научных трудов. – М.: Ин-т психологии РАН, 2010. – С. 135–152.

85. Кибрик А. А. Язык как он есть: русский мультимодальный дискурс: карточка проекта [Электронный ресурс] // А. А. Кибрик. – Ин-т языкознания РАН. – Электрон. дан. – Москва, 2016. Режим доступа: http://rscf.ru/prjcard_int?14-18-03819.

86. Кибрик А. Е. Эскиз лингвистической модели текстообразования / А. Е. Кибрик // Очерки по общим и прикладным вопросам языкознания (универсальное, типовое и специфичное в языке). – М., 1992. – С. 287–301.

87. Киселева С. В., Смирнова А. А., Трофимова Н. А. «Чат-бот коммуникация» как объект лингвистического исследования в системе цифровых коммуникаций / С. В. Киселева, А. А. Смирнова, Н. А. Трофимова // Дискурс. – 2022. – Т. 8, № 3. – С. 128–146.

88. Козориз А. В. Чат-боты как новый инструмент организации взаимодействия с клиентом / А. В. Козориз // Экономика: вчера, сегодня, завтра. – 2019. – Том 9. – №10А. – С. 639–648.

89. Колокольцева Т. Н. Проблемы изучения диалогической коммуникации и специфика интернет-диалогов (на материале блога) / Т. Н. Колокольцева // Культура и текст. – 2021. – №1 (44). – С. 197–215.

90. Колокольцева Т. Н. Диалогичность в жанрах интернет-коммуникации (чат, форум, блог) / Т. Н. Колокольцева // Жанры речи. Междунар. науч. журнал. – 2016. № 2. С. 96–104.

91. Коновалова М. В. Медиадискурс и подходы к его изучению / М. В. Коновалова // Вестник Челябинского государственного университета. – 2015. – №27 (382). – С. 101–107.
92. Кононова И. В., Николаев Ф. А. Seo-оптимизированный текст как жанровая разновидность компьютерно-опосредованного рекламного дискурса / И. В. Кононова, Ф. А. Николаев // Вестник Волгоградского государственного университета. Серия 2: Языкознание. – 2016. – №4. – С. 96–103.
93. Копытова А. В. Лингвопрагматические особенности ситуации общения «человек – чат-бот» / А. В. Копытова // Человек: Образ и сущность. Гуманитарные аспекты. – 2023. – №2 (54). – С. 123–139.
94. Креолизованный текст: Смысловое восприятие. Коллективная монография / Отв. ред. И.В. Вашунина. Ред. колл.: Е. Ф. Тарасов, А. А. Нистратов, М. О. Матвеев. – М.: Институт языкознания РАН, 2020. – 206 с.
95. Кубрякова Е. С. Эволюция лингвистических идей во второй половине XX века (опыт парадигмального анализа) / Е. С. Кубрякова // Язык и наука конца 20 века. – М., 1995. – С. 208–227.
96. Курилович Е. Очерки по лингвистике / Е. Курилович. – Биробиджан: ИП «ТРИВИУМ», 2000. – 490 с.
97. Латипова А. Л. Лингвистические особенности языка интернет-дискурса / А. Л. Латипова // Известия Волгоградского государственного педагогического университета. – 2018. – №6 (129). – С. 159–163.
98. Латур Б. Пересборка социального: введение в акторно-сетевую теорию / Б. Латур. – М.: Изд. дом Высшей школы экономики, 2014. – 384 с.
99. Лебедева М. Ю., Веселовская Т. С., Купрещенко О. Ф. Особенности восприятия и понимания цифровых текстов: междисциплинарный взгляд / М. Ю. Лебедева, Т. С. Веселовская, О. Ф. Купрещенко // Перспективы науки и образования. – 2020. – №4 (46). – С. 74–98.
100. Лингвистический процессор для сложных информационных систем / Ю. Д. Апресян, И. М. Богуславский, Л. Л. Иомдин и др.; Отв. ред. Л. П. Крысин; Институт проблем передачи информации (Москва). – М.: Наука, 1992. – 256 с.

101. Ляшевская О. Н. Корпусные инструменты в грамматических исследованиях русского языка / О. Н. Ляшевская. – М.: Издательский дом ЯСК, Рукописные памятники Древней Руси, 2016. – 520 с.
102. Манович Л. Язык новых медиа / Л. Манович; перевод Дианы Кульчицкой. – М.: Ад Маргинем Пресс, 2018. – 399 с.
103. Маринова Е. В. Оппозиция «человек – машина» в языковой картине мира XXI века / Е. В. Маринова // Гуманитарный вектор. – 2022. – Т. 17, №2. – С. 161–170.
104. Медова А. А. Онтология модальности: дисс. д-ра философ. наук: 09.00.01 / Медова Анастасия Анатольевна, 2016. – 303 с.
105. Мельник Н.В. Деривационное функционирование русского текста: лингвоцентрический и персонцентрический аспекты. – Кемерово, 2011. – 403 с.
106. Мельчук И. А. Опыт теории лингвистических моделей «Смысл-Текст»: семантика, синтаксис / И. А. Мельчук. – М.: Языки русской культуры, 1999. – 371 с.
107. Мельчук И. А. Язык: от смысла к тексту / И. А. Мельчук. — М.: Языки славянских культур, 2012. — 176 с.
108. Методы когнитивного анализа семантики слова: компьютерно-корпусный подход / Под общ. ред. В. И. Заботкиной. – 2-е изд. – М.: Издательский Дом ЯСК, 2019. – 344 с.
109. Мурзин Л. Н. Деривация в синхронном и диахронном аспектах / Л. Н. Мурзин // Деривация и история языка: межвуз. сб. науч. тр. / Перм. гос. ун-т. – Пермь, 1987. С. 4–10.
110. Мурзин Л. Н. О деривационных механизмах текстообразования / Л. Н. Мурзин // Теоретические аспекты деривации: межвуз. сб. науч. тр. / Перм. гос. ун-т. – Пермь, 1982. – С. 20–29.
111. Мурзин Л. Н. Основы дериватологии / Л. Н. Мурзин. – Пермь: Перм. ун-т, 1984. – 56 с.

112. Мурзин Л. Н. Синтаксическая деривация. Анализ производных предложений русского языка: пособие по спецкурсу / Л. Н. Мурзин. – Пермь: Перм. ун-т, 1974. – 128 с.
113. Национальный корпус русского языка: 2003–2005 результаты и перспективы, [сборник] / Рос. акад. наук, Ин-т рус. яз. им. В. В. Виноградова. – Москва: Индрик, 2005. – 343 с.
114. Омеляненко В. А. Поликодовые тексты в аспекте теории мультимодальности / В. А. Омеляненко // Коммуникативные исследования. – 2018. – №3. – С. 66–78.
115. Остапенко С. В. Аксиологическая и коммуникативная трансформация текста в парадигме цифровой реальности / С. В. Остапенко // Вестник Бурятского государственного университета. Филология. – 2024. – Вып. 4. – С. 3–11.
116. Остапенко С. В. Коммуникативная модель сообщения в процессе взаимодействия пользователя и диалогового агента в системах генерации текста / С. В. Остапенко // Научные исследования и разработки. Современная коммуникативистика. – 2025. – Т. 15. – №2 (75). – С. 79–85.
117. Остапенко С. В., Халина Н. В. Лингвистическая оптимизация промта в аспекте проблематики промт-инжиниринга / С. В. Остапенко, Н. В. Халина // Вестник Тюменского государственного университета. Гуманитарные исследования. Humanitates. – 2025. – Том 11. – №1 (41). – С. 36–50.
118. Остин Дж. Как совершать действия при помощи слов // Дж. Остин / Избранное / пер. с англ. Л. Б. Макеевой, В. П. Руднева. – М.: Идея Пресс, Дом интеллектуальной книги, 1999. – С. 13–136.
119. Пильгун М. А. Когнитивная модель анализа контента цифровых коммуникаций = Cognitive model of content analysis of digital communications / М. А. Пильгун // Вопросы психолингвистики. – 2023. – №2 (56). – С. 68–93.
120. Пищальникова В. А. Определение цифрового текста как методологическая проблема / В. А. Пищальникова // Вестник Московского государственного лингвистического университета. Гуманитарные науки. – 2022. – Вып. 6 (861). – С. 9–14.

121. Пищальникова В. А. Сущностные признаки цифрового текста / В. А. Пищальникова // Вестник Пятигорского государственного университета. – 2021. – Вып. 2. – С. 12–125.
122. Плотникова, А. М. Нейросеть как ключевое слово текущего момента / А. М. Плотникова // Филологический класс. – 2023. – Т. 28, № 2. – С. 45–54.
123. Плунгян В. А. Корпус как инструмент и как идеология: о некоторых уроках современной корпусной лингвистики / В. А. Плунгян // Русский язык в научном освещении. – 2008. – №16 (2). – С. 7–20.
124. Плунгян В. А. Корпусная лингвистика на современном этапе / В. А. Плунгян // Вестник Российской академии наук. – 2024. – Т. 94. – № 9. – С. 787–794.
125. Плунгян В. А. О перспективах современной корпусной лингвистики / В. А. Плунгян // В. А. Тишков (ред.). Труды Отделения историко-филологических наук РАН. – М.: Наука, 2016. – С.128-132.
126. Полоян А. В., Приходько К. Д. Составление промпта в нейросети midjourney для генерации иллюстраций к творчеству А. П. Чехова / А. В. Полоян, К. Д. Приходько // Информационные технологии в гуманитарных исследованиях: сборник трудов конференции. Красноярск, 25–28 сентября 2023 года. – Красноярск: Сибирский федеральный университет, 2023. – С. 194–199.
127. Понкин Д. И. Концепт предобученных языковых моделей в контексте инженерии знаний / Д. И. Понкин // International Journal of Open Information Technologies. – 2020. – Т. 8. – №. 9. – С. 18–29.
128. Потапова Р. К., Курьянова И. В. Особенности исследования текста в эпоху цифровой коммуникации / Р. К. Потапова, И. В. Курьянова // Вестник Волгоградского государственного университета. Серия 2, Языкознание. – 2021. – Т. 20, № 2. – С. 5–15.
129. Потемкин С. Б. Исследование оценочной функции высказывания с использованием нейронных сетей / С. Б. Потемкин // Международный журнал гуманитарных и естественных наук. – 2020. – №. 9–2. – С. 129–133.

130. Прошина М. В. Современные методы обработки естественного языка: нейронные сети / М. В. Прошина // Экономика строительства. – 2022. – № 5. – С. 27–42.
131. Разина И. Г. Деривационные механизмы текстопорождения: семантика, синтактика, прагматика / И. Г. Разина. – М.: Lap Lambert Academic Publishing, 2012. – 242 с.
132. Райскина В. А., Дубнякова О. А. Современные методы корпусной лингвистики при анализе текста (на примере корпуса BFM) / В. А. Райскина, О. А. Дубнякова // Актуальные вопросы современной науки. – 2015. – №40. – С. 146–154.
133. Резанова З. И. Внутренняя форма слова как объект метаязыковой рефлексии в условиях чат-коммуникации / З. И. Резанова // Язык и культура. – 2008. – №1. – С. 78–85.
134. Рудак Л. В., Федяев О. И. Анализ подходов к решению проблемы понимания и обработки естественного языка методами машинного обучения / Л. В. Рудак, О. И. Федяев // Современные информационные технологии в образовании и научных исследованиях (СИТОНИ-2021). – 2021. – С. 216–224.
135. Рябцева Н. К. Отличительные особенности интернет-коммуникации в металингвистическом и прикладном аспекте / Н. К. Рябцева // Лингвистика и методика преподавания иностранных языков». Выпуск 9. – 2017. – С. 387–429.
136. Савчук С. О., Архангельский Т. А., Бонч-Осмоловская А. А. Национальный корпус русского языка 2.0: новые возможности и перспективы развития / С. О. Савчук, Т. А. Архангельский, А. А. Бонч-Осмоловская // Вопросы языкознания. – 2024. – №2. – С. 7–34.
137. Садур В. Г. Речевое общение с электронно-вычислительными машинами и проблемы их развития / В. Г. Садур // Речевое общение: проблемы и перспективы / Отв. ред. Ф. М. Березин. – Москва: ИНИОН АН СССР, 1983. – С. 78–113.
138. Северина Е. М. Цифровой текст в пространстве современной культуры / Е. М. Северина // Человек и культура. – 2019. – №5. – С. 65–72.

139. Сериков А. Е. Проблема порождения смысла и современный генеративизм / А. Е. Сериков // Вестник Самарской гуманитарной академии. Сер.: Философия. Филология. – 2007. – №2. – С.155–173.
140. Симанков В. С., Теплоухов С. В. Аналитическое исследование методов и алгоритмов искусственного интеллекта / В. С. Симанков, С. В. Теплоухов // Вестник Адыгейского государственного университета. Серия 4: Естественно-математические и технические науки. – 2020. – №3 (266). – С. 16–25.
141. Скоринкин Д. А. Электронное представление текста с помощью стандарта разметки TEI / Д. А. Скоринкин // Вестник Московского университета. Серия 9. Филология. – 2016. №5. – С. 90–108.
142. Скоринкин Д. А., Бонч-Осмоловская А. А. «Особые приметы» в речи художественных персонажей: количественный анализ диалогов в «Войне и мире» Л. Н. Толстого [Электронный ресурс] / Д. А. Скоринкин, А. А. Бонч-Осмоловская // Электронный научно-образовательный журнал «История». – 2016. – Т. 7. – № 7 (51). Режим доступа: <http://history.jes.su/s207987840001649-2-1>.
143. Скоринкин Д. А. Семантическая разметка художественных текстов для количественных исследований в филологии (на примере романа «Война и мир» Л. Н. Толстого): дис. ... канд. филол. наук / Скоринкин Даниил Андреевич, 2018. – 164 с.
144. Скороходов И. С., Тихомирова А. Н. Основные этапы обработки текста и генерации признаков в задачах текстовой классификации / И. С. Скороходов, А. Н. Тихомирова // Проблемы Науки. – 2016. – №15 (57). – С. 18–22.
145. Соколова Е. Г., Шаров С. А. К многоязыковой генерации руководств пользователя: начальный этап проекта AGILE / Е. Г. Соколова, С. А. Шаров // Труды Международного семинара «Диалог'98» по компьютерной лингвистике и ее приложениям (ред. А. С. Нариньяни). – Казань, 1998. – С. 848–859.
146. Соколова Е. Г., Шаров С.А. Многоязыковая генерация руководств пользователя на славянских языках / Е. Г. Соколова, С. А. Шаров // КИИ'2000, Т.1. – М.: Изд-во физ-мат. лит-ры, 2000. – С. 409–417

147. Сорокин Ю. А. Креолизованные тексты и их коммуникативная функция / Ю. А. Сорокин, Е. Ф. Тарасов // Оптимизация речевого воздействия. – Москва: Высшая школа, 1990. – С. 180–186.
148. Соснин П. И. Вопросно-ответное программирование человеко-компьютерной деятельности / П. И. Соснин. – Ульяновск: УлГТУ, 2010. – 240 с.
149. Тельпов Р. Е., Ларцина С. В. Типовые различия естественных и сгенерированных нейронной сетью текстов в квантитативном аспекте / Р. Е. Тельпов, С. В. Ларцина // Научный диалог. – 2023. – Т. 12. – № 7. – С. 47–65.
150. Тестелец Я. Г. Глава XI. Порождающая грамматика: от правил к ограничениям / Я. Г. Тестелец // Введение в общий синтаксис. – М.: РГГУ. – 2001. – 800 с.
151. Тичер С., Мейер М., Водак Р. Методы анализа текста и дискурса / С. Тичер, М. Мейер, Р. Водак. – Харьков, 2009. – 356 с.
152. Трубникова Ю. В. Лексико-деривационные основания моделирования текста : монография / Ю. В. Трубникова ; Федеральное агентство по образованию России, Алтайский гос. ун-т. – Барнаул : Изд-во Алтайского гос. ун-та, 2008. – 179 с.
153. Тюрина Д. А., Пальмов С. В. Применение нейронных сетей в обработке естественного языка / Д. А. Тюрина, С. В. Пальмов // Журнал прикладных исследований. – 2023. – №7. – С. 158–162.
154. Филатова О. Г. Дискурсивные практики в цифровой публичной сфере современной России: результаты и направления исследований / О. Г. Филатова // Управленческое консультирование. – 2019. – №11. – С. 180–192.
155. Халина Н. В., Злобина Ю. И., Авдеева Т. Н. и др. Лингвистика инжиниринга: русский язык как система самоописания общества / Н. В. Халина, Ю. И. Злобина, Т. Н. Авдеева, Т. С. Хребтова, Н. Н. Столярова, В. С. Белоусова. – Барнаул: Изд-во АлтГУ, 2013. – 234 с.
156. Хутыз И. П. Мультиmodalность академического дискурса как условие его коммуникативной успешности / И. П. Хутыз // Вестник Адыгейского

государственного университета. Серия 2: Филология и искусствоведение. – 2016. – №1. – С. 90–95.

157. Цифровые гуманитарные исследования: монография / А. Б. Антопольский, А. А. Бонч-Осмоловская, Л. И. Бородкин [и др.]. – Красноярск: Сиб. федер. ун-т, 2023. – 272 с.

158. Чейф У. Память и вербализация прошлого опыта / У. Чейф // Новое в зарубежной лингвистике: Прикладная лингвистика. – М., 1983. Вып. XII. С. 35–74.

159. Чемодурова З. М. Искусство повествования в цифровую эпоху: мультимодальный аспект / З. М. Чемодурова // Вестник Волгоградского государственного университета. Серия 2, Языкознание. – 2022. – Т. 21, № 6. – С. 110–120. – (На англ. яз.).

160. Черкасова М. Н., Тактарова А. В. Искусственно сгенерированный академический текст (лингвопрагматический аспект) / М. Н. Черкасова, А. В. Тактарова // Филологические науки. Вопросы теории и практики. – 2024. – №7. – С. 2551–2557.

161. Черкасова М. Н., Тактарова А. В. Признаки сгенерированного текста в академическом дискурсе: проблема идентификации / М. Н. Черкасова, А. В. Тактарова // Филологические науки. Вопросы теории и практики. – 2024. – №7. – С. 2226–2232.

162. Черненькая С. В. Текст в цифровом пространстве культуры / С. В. Черненькая // Вестник Московского городского педагогического университета. Серия: Философские науки. – 2018. – №3 (27). – С. 58–63.

163. Чернышова Т. В. Филологическое исследование медиакommunikации: теоретико-методологические предпосылки и методы анализа текста и дискурса (ретроспектива) / Т. В. Чернышова // Медиалингвистика. – 2014. – №1 (4). – С. 77–86.

164. Чувакин А. А. Деривационные отношения как тип межтекстовых отношений (к предмету текстодериватологии) / А. А. Чувакин // Актуальные проблемы дериватологии, мотивологии, лексикографии. – Томск: Изд-во Томского ун-та, 1998. – С. 23–24.

165. Шаумян С. К. Структурная лингвистика / С. К. Шаумян. – М.: «Наука», 1965. – 396 с.
166. Шведова Н. Ю. Очерки по синтаксису русской разговорной речи / Н. Ю. Шведова. – Москва: изд-во Академии наук СССР, 1960. – 377 с.
167. Шерстинова Т. Ю. Наиболее употребительные слова повседневной русской речи (в гендерном аспекте и в зависимости от условий коммуникации) / Т. Ю. Шерстинова // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (Москва, 1–4 июня 2016 г.). Вып. 15 (22). – М.: Изд-во РГГУ, 2016. – С. 616–631.
168. Шерстинова Т. Ю. Прагматическое аннотирование коммуникативных единиц в корпусе ОРД: микроэпизоды и речевые акты / Т. Ю. Шерстинова; ред. Захаров В. П., Хохлова М. В. // Труды международной конференции «Корпусная лингвистика – 2015». – СПб.: СПбГУ. – С. 436–445.
169. Шерстинова Т. Ю. Структура повседневного диалога как последовательность речевых актов / Т. Ю. Шерстинова // Компьютерная лингвистика и интеллектуальные технологии: по материалам международной конференции «Диалог 2018» (Москва, 30 мая — 2 июня 2018 г.). Вып. 17 (24) / гл. ред. В. П. Селегей. – М.: РГГУ, 2018. – С. 637–651.
170. Шилина М. Г. Текстогенность как фактор трансформации коммуникации в XXI веке (к вопросу формирования инновационных концепций исследований медиа и PR) / М. Г. Шилина // Вестник Московского университета. Серия 10. Журналистика. – 2012. – №5. – С. 108–117.
171. Шмаков А. А. Коммуникативное моделирование как методика изучения бытия текста с обращением в интернет-коммуникации / А. А. Шмаков // Мир науки, культуры, образования. – 2013. – №4 (41). – С. 15–17.
172. Шпильная Н. Н. Принцип суппозиционной связи лексемы и текста как ключевой принцип организации человеко-компьютерной коммуникации / Н. Н. Шпильная // Культура и текст. – 2018. – №4 (35). – С. 209–226.
173. Шпильная Н. Н. Диалогический текст: деривационная концепция, [монография] / Н. Н. Шпильная. – Москва: URSS, ЛЕНАНД, 2018. – 383 с.

174. Шпильная Н.Н. Внутренняя форма текста как деривационный феномен / Н. Н. Шпильная // Вестник Томского государственного университета. – 2013. – №373. – С. 44–50.
175. Шпильная Н.Н. Внутренняя форма текста как носитель потенциала его деривационного развития / Н. Н. Шпильная // Вестник Кемеровского государственного университета. – 2012. – № 4 (52). – С. 232–235.
176. Шульман В. Д., Максименко В. Д., Волхонцева П. Д. Анализ программных средств морфологического анализа / В. Д. Шульман, В. Д. Масименко, П. Д. Волхонцева // Международный журнал гуманитарных и естественных наук. – 2022. – №3–2. – С. 166–170.
177. Щипицина Л. Ю. Компьютерно-опосредованная коммуникация. Лингвистический аспект анализа / Л. Ю. Щипицина. – Москва: Красанд, 2010. – 294 с.
178. Яцко В. А. Автоматическое распознавание жанра и адаптивное реферирование текста / В. А. Яцко // Научно-техническая информация. Сер. 2, Информационные процессы и системы. – 2010. – №5. – С. 9–18.
179. Яцко В. А. Методы и алгоритмы автоматического анализа текста / В. А. Яцко // Научно-техническая информация. Сер. 2, Информационные процессы и системы. – 2011. – №9. – С. 12–19.
180. Яцко В. А. Проблемы информационно-лингвистического моделирования научного текста: диссертация ... доктора филологических наук: 05.13.17. / Яцко Вячеслав Александрович, 1998. – 341 с.
181. Adami E., Kress G. Introduction: Multimodality, Meaning Making, and the Issue of “Text” / E. Adami, G. Kress // Text & Talk. – 2014. – Vol. 34. Iss. 3. – Pp. 231-243.
182. Baker Mark C. The Atoms of Language: The Mind's Hidden Rules of Grammar / Mark C Baker. – Publisher. New York: Basic Books, 2001. – 276 p.
183. Barthes R. Elements of semiology. Annette Lavers & Colin Smith (Trans.) / R. Barthes. – New York: Hill and Wang, 1977. – 112 p.

184. Barthes R. *The empire of signs* / R. Barthes. – New York: Hill and Wang, 1982. –128 p.
185. Bird S., Klein E., Loper E. *Natural Language Processing with Python* / S. Bird, E. Klein, E. Loper. – O'Reilly Media, 2009. – 482 p.
186. Chomsky N. *New Horizons in the Study of Language and Mind* / N. Chomsky. – Cambridge, 2001. – 256 p.
187. Chomsky N. *Remarks on nominalizations. Readings in English transformational grammar* / N. Chomsky. – Waltham (Mass.) etc.: Ginn, 1970. – Pp. 184–221.
188. Chomsky N. *Syntactic Structures* / N. Chomsky. – S'-Gravenhage, 1957. – 116 p.
189. Chomsky N. *The sound pattern of English* / N. Chomsky, M. Halle. – N. Y.; Evanston; L., 1968. – 470 p.
190. Chomsky N. *A Review of B. F. Skinner's Verbal Behavior* / N. Chomsky // *Language*. – 1959. – Vol. 35, No. 1. – Pp. 26–58.
191. Clark A., Fox C., Lappin S. *The Handbook of Computational Linguistics and Natural Language Processing* / A. Clark, C. Fox, S. Lappin. – Wiley-Blackwell, 2013. – 800 p.
192. Crystal D. *Internet Linguistics: A Student Guide* / D. Crystal. – NY.: Routledge, 2011. – 172 p.
193. Deemter K. van, Kraemer E. *Graphs and Natural Language Processing* / K. van Deemter, E. Kraemer. — Springer, 2015. – 240 p.
194. *Discourse and digital practices: Doing Discourse Analysis in the Digital Era*. Eds. R.H. Jones, A. Chik, C.A. Hafner. L. – N.Y.: Routledge, 2015. XII. – 250 p.
195. Dredze M., Crammer K. *Online Methods for Multi-Domain Learning* / M. Dredze, K. Crammer. – EMNLP, 2008. – Pp. 689-697.
196. Eigner F. S. *Combinatory Categorical Grammar: Tutorial and Survey* / F. S. Eigner. – 2016. – arXiv:1603.04719.
197. Eisner J. *Parameter Estimation for Probabilistic Finite-State Transducers* / J. Eisner. – ACL, 2002. – Pp. 1–8.

198. Francis W.N., Kucera H. Frequency Analysis of English Usage: Lexicon and Grammar / W. N.Francis, H. Kucera. – Boston: Houghton Mifflin, 1982. – 710 p.
199. Gibbs J. L., Kirkwood G. L., Fang C. & Wilkenfeld J. N. Negotiating agency and control: Theorizing human-machine communication from a structurational perspective / J. L. Gibbs, G. L. Kirkwood, C. Fang & J. N. Wilkenfeld // Human-Machine Communication. – 2021. – No.2. – Pp. 153-171.
200. González-Arias C., Chatzikoumi E. & López-García X. The Anthropomorphic Pursuit of AI-Generated Journalistic Texts: Limits to Expressing Subjectivity / C. González-Arias, E. Chatzikoumi & X. López-García// Frontiers in Communication. – 2024. – No. 9. – Available at: https://www.researchgate.net/publication/384965498_The_anthropomorphic_pursuit_of_AI-generated_journalistic_texts_limits_to_expressing_subjectivity
201. Grice H. P. Meaning / H. P. Grice // Philosophical Rev. – 1957. – Vol. 66, № 3. – Pp. 377–388.
202. Hale K. & Jay K. The basic elements of argument structure / K. Hale & K. Jay // Papers from the UPenn/MIT Roundtable on Argument Structure and Aspect, ed. by Heidi Harley. Held at MIT in spring 1997. – MIT Working Papers in Linguistics. – Vol. 32. Cambridge, Mass.: MIT Working Papers in Linguistics. 1998. – Pp.73-118.
203. Halliday M. A. K. Language as Social Semiotic: The Social Interpretation of Language and Meaning / M. A. K. Halliday. – University Park Press, 1978. – 192 p.
204. Jackendoff R. A Parallel Architecture perspective on language processing / R. Jackendoff // Brain Research. – 2007. – Pp. 2–22.
205. Jackendoff R. Foundations of Language: Brain, Meaning, Grammar, Evolution / R. Jackendoff. – Oxford University Press, 2002. – 477 p.
206. Jewitt C. An Introduction to Multimodality / C. Jewitt // The Routledge Handbook of Multimodal Analysis. – London; New York: Routledge, 2009. – Pp.14–27.
207. Jones S., Gunkel D. J., Edwards A. P., etc. Human-Machine Communication: Rethinking Communication, Technology, and Ourselves / S. Jones, D. J. Gunkel, A. P. Edwards, P. R. Spence, E. Chad. – Peter Lang Verlag, 2018. – 179 p.

208. Katz J. J. An integrated theory of linguistic descriptions / J. J. Katz, P. M. Postal. – Cambridge; Mass: MIT Press, 1964. – 178 p.
209. Koehn P. Statistical Machine Translation / P. Koehn. – Cambridge University Press, 2010. – 433 p.
210. Kolmogorova A. V, Margolina A. V. Written vs generated text: “naturalness” as a textual and psycholinguistic category / A. V. Kolmogorova, A. V. Margolina // Научный результат. Вопросы теоретической и прикладной лингвистики. – 2024. – №2. – С. 71-99.
211. Kress G. Multimodal Discourse: The modes and media of contemporary communication / G. Kress. – London: Edward Arnold, 2001. – 152 p.
212. Kress G. Multimodality. A Social Semiotic Approach to Contemporary Communication / G. Kress. – London: Routledge, 2010. – 232 p.
213. Kress G. What Is Mode? / G. Kress // A Handbook of Multimodal Analysis. – N. Y.: Routledge, 2009. – Pp. 54-67.
214. Kress G., van Leeuwen T. Multimodal discourse. The modes and media of contemporary communication / G. Kress, T. van Leeuwen. – London: Arnold, 2001. – 142 p.
215. Lasnik H. & Saito M. Move Alpha: Conditions on Its Application and Output / H. Lasnik & M. Saito. – MIT Press, 1992. – 234 p.
216. Lasnik H. Current studies in linguistics series. A course in GB syntax: Lectures on binding and empty categories / H. Lasnik. – Cambridge (Mass.), London: The MIT press, The MIT press, 1988. XII. – 178 p.
217. Leeuwen T. van. Introducing social semiotics / T. van Leeuwen. – L.: Routledge, 2005. – 314 p.
218. Manning C. D., Schütze H. Foundations of Statistical Natural Language Processing / C. D. Manning, H. Schütze. – MIT Press, 1999. – 680 p.
219. Murray J. Composing Multimodality / J. Murray // Multimodal Composition: A Critical Sourcebook. – Boston: Bedford; St. Martin's, 2013. – 544 p.
220. Neubig G. Neural Machine Translation / G. Neubig. – MIT Press, 2021. – 300 p.

221. O'Halloran K. L. Inter-Semiotic Expansion of Experiential Meaning: Hierarchical Scales and Metaphor in Mathematics Discourse / K. L. O'Halloran // *New Developments in the Study of Ideational Meaning: From Language to Multimodality*. – London: Equinox, 2008. – Pp. 231-254.
222. O'Toole M. The language of displayed art / M. O'Toole. – London: Leicester university press, 1994. XIII. – 295 p.
223. Pollock J.-Y. Verb Movement, UG and the Structure of IP / J.-Y. Pollock // *Linguistic Inquiry*. – 1989. – No. 20(3). – Pp. 365-424.
224. Rizzi L. Relativized Minimality / L. Rizzi. – Publisher: The MIT Press, 1990. – 160 p.
225. Scollon R. Mediated Discourse: The Nexus of Practice / R. Scollon. – Routledge, 2001. – 196 p.
226. Searle Jo. What is a speech act? / Jo. Searle // *Philosophy in America* / ed. M. Black. – London: Alien and Unwin, 1965. – Pp. 221–239.
227. Shaumyan S. A. Semiotic Theory of Language / S. A. Shaumyan. – Indiana University Press, 1987. – 372 p.

Результаты экспериментальной генерации текстов на основе параметризации промтов

Текст 1. Онлайн-обучение – это удобный и гибкий способ приобретения новых навыков и знаний. С интернетом у нас под рукой, мы можем получить доступ к огромному количеству курсов и образовательных ресурсов из любой точки мира. Этот способ обучения позволяет нам учиться в своем темпе, в любое время и в любом месте. Онлайн-обучение особенно полезно для тех, кто имеет плотный график или не может посещать традиционные классы.

Текст 2. Онлайн-обучение имеет несколько преимуществ перед традиционным обучением. Оно обеспечивает гибкость, позволяя студентам учиться в любое время и в любом месте, что особенно важно для тех, кто имеет ограничения в графике. Кроме того, онлайн-обучение обеспечивает доступность, позволяя студентам из различных регионов и стран получать доступ к качественному образованию. Это особенно важно для тех, кто живет в регионах с ограниченными возможностями доступа к образованию.

Текст 3. Онлайн-обучение может быть особенно полезным для учащихся с ограниченными возможностями, обеспечивая им инклюзивность и равные возможности в получении образования. Онлайн-курсы могут быть адаптированы к индивидуальным потребностям каждого студента, что позволяет им учиться в комфортной и доступной среде. Кроме того, онлайн-обучение обеспечивает доступность образования для учащихся с ограниченными возможностями, которые могут иметь трудности с посещением традиционных классов.

Преимущества онлайн-обучения для учащихся с ограниченными возможностями:

Инклюзивность: онлайн-обучение обеспечивает равные возможности для всех студентов, независимо от их физических или умственных ограничений.

Доступность: онлайн-обучение позволяет студентам с ограниченными возможностями получать доступ к качественному образованию, которое может быть недоступно им в традиционных классах.

Гибкость: онлайн-обучение позволяет студентам учиться в своем темпе, что особенно важно для тех, кто имеет ограничения в графике.

Проблемы онлайн-обучения для учащихся с ограниченными возможностями:

Технические проблемы: онлайн-курсы могут требовать специального оборудования или программного обеспечения, которое может быть недоступно для некоторых студентов.

Ограничения в доступе: некоторые студенты с ограниченными возможностями могут иметь ограничения в доступе к интернету или компьютеру, что может препятствовать их участию в онлайн-обучении.

Текст 4. Простой: Влияние фейковых новостей на общество

Каково влияние фейковых новостей на общество? Фейковые новости могут иметь серьезные последствия для общества, включая дезинформацию граждан, подрыв доверия к СМИ и политическим институтам, а также провоцирование социальной напряженности и конфликтов. Они могут также влиять на наши решения и выборы, основанные на ложной информации. Влияние фейковых новостей на общество – это серьезная проблема, требующая внимания и решительных действий.

Текст 5. Фейковые новости имеют значительное влияние на общественное мнение, подрывая доверие к СМИ и политическим институтам. Дезинформация, распространяемая через фейковые новости, может влиять на наши убеждения и мнения, основанные на ложной информации. Это может привести к формированию неправильных представлений о политических событиях, социальных проблемах и экономических тенденциях. Поэтому особенно важно обеспечивать достоверность информации, чтобы предотвратить распространение фейковых новостей и защитить общественное мнение от дезинформации.

Текст 6. Социальные сети играют значительную роль в распространении фейковых новостей, что может иметь серьезные последствия для общества. Алгоритмы социальных сетей могут способствовать распространению фейковых новостей, делая их более привлекательными и убедительными для пользователей. Кроме того, социальные сети могут создавать "пузыри фильтрации", где пользователи видят только информацию, подтверждающую их убеждения, и не видят альтернативных точек зрения. Проверка фактов и критическое мышление становятся особенно важными в борьбе с фейковыми новостями в социальных сетях.

Причины распространения фейковых новостей в социальных сетях:

Алгоритмы социальных сетей, которые могут способствовать распространению фейковых новостей.

"Пузыри фильтрации", которые ограничивают доступ к альтернативной информации.

Недостаток критического мышления и проверки фактов у пользователей социальных сетей.

Следствия распространения фейковых новостей в социальных сетях:

Дезинформация граждан и подрыв доверия к СМИ и политическим институтам.

Формирование неправильных представлений о политических событиях, социальных проблемах и экономических тенденциях.

Провоцирование социальной напряженности и конфликтов на основе ложной информации.

Текст 7. Круговорот воды в природе, также называемый водным циклом, представляет собой непрерывный процесс испарения, конденсации, дождя и течи, обеспечивающий постоянное движение воды между землей, атмосферой и растительностью. Вода испаряется с поверхности океанов, морей, озёр и других водоемов, конденсируется в виде облаков и, в конечном итоге, возвращается на землю в виде осадков. Таким образом, круговорот воды поддерживает жизнь на планете, обеспечивая растения, животных и людей доступом к воде.

Текст 8. Круговорот воды играет решающую роль в поддержании экосистем, поскольку он обеспечивает постоянное движение воды между атмосферой, гидросферой и геосферой. Ключевые процессы в круговороте воды, такие как испарение и конденсация, помогают регулировать температуру планеты, формировать погодные условия и поддерживать жизнь в экосистемах. Испарение, например, способствует охлаждению поверхности земли, в то время как конденсация обеспечивает осадки, необходимые для роста растений и поддержания водных ресурсов. Таким образом, круговорот воды имеет решающее значение для поддержания баланса экосистем и жизни на планете.

Текст 9. Изменение климата, вызванное глобальным потеплением, оказывает значительное влияние на водный цикл, изменяя причины и следствия круговорота воды. Глобальное потепление приводит к увеличению температуры воздуха, что, в свою очередь, увеличивает испарение и конденсацию. Это может привести к изменению погодных условий, таких как более частые засухи и наводнения, а также к изменению распределения осадков по всему миру. Кроме того, таяние ледников и снега в результате глобального потепления может привести к повышению уровня моря и затоплению прибрежных районов. Таким образом, изменение климата может иметь серьезные последствия для водного цикла и жизни на планете, подчеркивая необходимость принятия мер по снижению выбросов парниковых газов и сохранению водных ресурсов.

Текст 10. Роль наставника в образовании заключается в том, чтобы помочь ученикам достичь своих целей и развить свои навыки. Наставник играет важную роль в образовании, поскольку он обеспечивает руководство, поддержку и мотивацию ученикам, помогая им преодолевать трудности и достичь успеха. Наставник также помогает ученикам развивать свои навыки и знания, а также формирует у них положительное отношение к обучению.

Текст 11. Наставничество играет важную роль в образовании, поскольку оно обеспечивает руководство и поддержку ученикам. Ключевые преимущества наставничества включают в себя индивидуализированное руководство, которое помогает ученикам достичь своих целей, а также поддержку, которая мотивирует

их к обучению. Кроме того, наставничество помогает ученикам развивать свои навыки и знания, а также формирует у них положительное отношение к обучению. В целом, наставничество является важным инструментом в образовании, который помогает ученикам достичь успеха.

Текст 12. Наставничество играет важную роль в подготовке учителей, поскольку оно обеспечивает профессиональное развитие и коучинг. Профессиональное развитие учителей является ключевым фактором в обеспечении качества образования, и наставничество помогает учителям развивать свои навыки и знания. Коучинг, в свою очередь, помогает учителям преодолевать трудности и достичь успеха в своей работе. Преимущества наставничества в подготовке учителей включают в себя улучшение качества образования, увеличение мотивации учителей и развитие их навыков и знаний. Однако, наставничество также может иметь некоторые проблемы, такие как нехватка ресурсов и времени, а также необходимость в дополнительной поддержке для учителей. В целом, наставничество является важным инструментом в подготовке учителей, который помогает им достичь успеха и обеспечивать качество образования.

Текст 13. Телевидение имеет богатую историю, которая началась в конце 19 века. Первый телевизор был изобретен Полом Нипковым в 1884 году, но он не был способен передавать движущиеся изображения. В 1920-х годах Джон Логи Бэрд и Фил Фарнsworth независимо друг от друга разработали первую систему телевизионного вещания. В 1940-х годах телевидение стало популярным средством массовой информации, и к 1950-м годам оно стало основным источником развлечения и информации для миллионов людей. С тех пор телевидение продолжает развиваться, и сегодня оно доступно в различных форматах, включая цифровое телевидение, кабельное телевидение и интернет-телевидение.

Текст 14. Телевидение играет важную роль в обществе, обеспечивая развлечения и информацию для миллионов людей. Оно позволяет людям получать доступ к новостям, развлекательным программам и образовательным контентам

из любой точки мира. Телевидение также влияет на наше мировоззрение, формируя наше понимание событий и явлений. Однако, телевидение также может иметь негативное влияние, распространяя стереотипы и предвзятости. В целом, телевидение является мощным инструментом, который может быть использован для блага или зла, и его влияние на общество зависит от контента, который он транслирует.

Текст 15. Телевидение играет важную роль в формировании общественного мнения, и его влияние на наше мировоззрение не может быть переоценено. Телевидение может быть использовано как инструмент пропаганды, формируя общественное мнение в соответствии с интересами тех, кто контролирует его. Однако, телевидение также может быть использовано для информирования и просвещения людей, помогая им формировать свое мнение на основе фактов. В историческом контексте телевидение было использовано для пропаганды в период Второй мировой войны, а также в период холодной войны. Сегодня телевидение продолжает играть важную роль в формировании общественного мнения, и его влияние на наше мировоззрение остается значительным.

Текст 16. Фотосинтез – это процесс, при котором растения, водоросли и некоторые бактерии производят органические вещества из углекислого газа и воды, используя энергию солнечного света. Это происходит в хлоропластах, которые содержат пигмент хлорофилл, поглощающий световую энергию. Фотосинтез является основным процессом, который обеспечивает жизнь на Земле, так как он производит кислород и органические вещества, необходимые для роста и развития живых организмов.

Текст 17. Фотосинтез играет важную роль в росте растений, так как он обеспечивает им энергию и органические вещества, необходимые для развития. В этом процессе растения используют световую энергию для преобразования углекислого газа и воды в глюкозу и кислород. Глюкоза является основным источником энергии для растений, а кислород выделяется в атмосферу как побочный продукт. Фотосинтез также обеспечивает растения необходимыми

органическими веществами, такими как аминокислоты, сахара и другие соединения, которые необходимы для роста и развития.

Текст 18. Фотосинтез играет важную роль в экосистеме, так как он обеспечивает производство кислорода и органических веществ, необходимых для жизни живых организмов. В экосистеме фотосинтез является основным процессом, который поддерживает пищевую цепь. Растения, производящие органические вещества, становятся основой пищевой цепи, обеспечивая питанием животных и микроорганизмов. Кроме того, фотосинтез влияет на климат, так как он поглощает углекислый газ из атмосферы и выделяет кислород. В целом, фотосинтез является важным процессом, который поддерживает жизнь на Земле и играет ключевую роль в экосистеме.

Текст 19. Обратная связь играет ключевую роль в образовании, позволяя учащимся и преподавателям совершенствоваться и развиваться. Она позволяет получать информацию о результатах обучения, оценивать эффективность обучающих процессов и вносить необходимые изменения. Кроме того, обратная связь способствует мотивации учащихся, позволяя им осознавать свои успехи и слабые стороны, и, как следствие, улучшая качество обучения.

Текст 20. Обратная связь является неотъемлемым элементом эффективного обучения студентов. Она включает в себя оценивание результатов обучения и анализ эффективности обучающих процессов. Ключевые преимущества обратной связи в обучении студентов заключаются в следующем:

Помогает студентам осознавать свои успехи и слабые стороны, что способствует мотивации и улучшению качества обучения.

Позволяет преподавателям оценивать эффективность обучающих процессов и вносить необходимые изменения.

Позволяет идентифицировать пробелы в знаниях и навыках студентов, и, соответственно, корректировать обучающие процессы.

Содействует развитию критического мышления и самостоятельности у студентов.

Текст 21. Обратная связь играет важную роль в профессиональном развитии учителей, способствуя их рефлексии и росту. Ключевые преимущества и проблемы использования обратной связи в профессиональном развитии учителей заключаются в следующем:

Помогает учителям оценивать эффективность своих обучающих процессов и вносить необходимые изменения.

Позволяет идентифицировать пробелы в знаниях и навыках учителей, и, соответственно, корректировать их профессиональное развитие.

Содействует развитию критического мышления и самостоятельности у учителей.

Позволяет учителям получать обратную связь от коллег, экспертов и учащихся, что способствует их росту и развитию.

Проблемы:

Недостаточное понимание значения обратной связи в профессиональном развитии учителей.

Недостаточное количество времени и ресурсов, выделяемых на обеспечение обратной связи.

Недостаточная квалификация и подготовка учителей в области получения и использования обратной связи.

Недостаточная культура обратной связи в образовательных учреждениях.

В целом, обратная связь является важным инструментом в профессиональном развитии учителей, и ее использование должно быть активно поощряться и поддерживаться образовательными учреждениями.

Текст 22. Влияние социальных сетей на отношения является предметом активного обсуждения. С одной стороны, социальные сети могут способствовать укреплению связей между людьми, позволяя им общаться и поддерживать отношения на расстоянии. С другой стороны, они могут также негативно влиять на отношения, создавая иллюзию близости и подменяя личное общение виртуальным.

Текст 23. Социальные сети играют значительную роль в формировании коммуникационных процессов в современном обществе. Онлайн-взаимодействие в социальных сетях может иметь как положительные, так и отрицательные последствия для коммуникации. С одной стороны, социальные сети позволяют людям общаться и обмениваться информацией в режиме реального времени, что может способствовать укреплению связей и развитию эмоционального интеллекта. С другой стороны, они могут также создавать иллюзию близости и подменять личное общение виртуальным, что может негативно влиять на качество коммуникации.

Текст 24. Социальные сети играют важную роль в формировании идентичности в современном обществе. Самопрезентация в социальных сетях позволяет людям создавать онлайн-персону, которая может отличаться от их реальной личности. Это может привести к формированию ложной идентичности, основанной на виртуальном образе, а не на реальных качествах и характеристиках человека.

Причины:

Желание быть популярным и признанным в социальных сетях.

Необходимость соответствовать социальным нормам и ожиданиям.

Возможность создавать идеализированный образ себя.

Следствия:

Формирование ложной идентичности.

Негативное влияние на самооценку и самоуважение.

Утрата реальной связи с другими людьми и обществом в целом.

Текст 25. Человеческий мозг является сложным органом, состоящим из нескольких отделов, каждый из которых выполняет определенные функции.

Основные отделы мозга включают в себя:

Мозжечок: отвечает за координацию движений и равновесие.

Ствол мозга: регулирует автоматические функции организма, такие как дыхание и сердцебиение.

Просвет мозга: отвечает за обработку сенсорной информации и управление движениями.

Кортекс мозга: является центром высшей нервной деятельности, отвечает за мышление, память и эмоции.

Текст 26. Человеческий мозг является сложной системой, состоящей из миллиардов нейронов, которые взаимодействуют друг с другом через синапсы. Нейроны – это основные функциональные единицы мозга, которые передают и обрабатывают информацию. Синапсы – это места контакта между нейронами, где происходит передача сигналов.

Основные функции мозга включают в себя:

Обработку сенсорной информации.

Управление движениями.

Регулирование автоматических функций организма.

Мышление и память.

Эмоции и мотивация.

Текст 27. Развитие мозга играет важную роль в формировании когнитивных способностей. Пластичность мозга – это способность мозга изменяться и адаптироваться к новым условиям, что позволяет ему развивать новые связи и функции. Нейропластичность – это способность мозга изменять свою структуру и функцию в ответ на новые опыт и обучение.

Стадии развития мозга:

Эмбриональная стадия: формирование основных отделов мозга.

Период раннего развития: формирование нейронов и синапсов.

Период позднего развития: формирование сложных нейронных сетей и когнитивных функций.

Важность развития мозга для когнитивных способностей:

Формирование памяти и обучения.

Развитие языковых и математических способностей.

Формирование эмоциональной регуляции и социальных навыков.

Развитие критического мышления и решения проблем

Текст 28. Какова роль технологий в образовании? Технологии могут изменить способ, которым мы учимся и преподаем, но как они влияют на процесс образования?

Текст 29. Технологии могут значительно улучшить процесс образования, сделав его более доступным и персонализированным. Доступность технологий позволяет студентам учиться в любое время и в любом месте, что особенно важно для тех, кто имеет ограниченный доступ к образовательным ресурсам. Персонализация технологий позволяет преподавателям адаптировать образование к индивидуальным потребностям каждого студента, что может привести к более эффективному обучению.

Текст 30. Интеграция технологий в образование может иметь как преимущества, так и проблемы. С одной стороны, технологии могут:

Увеличить доступность образования для более широкой аудитории.

Обеспечить более персонализированное обучение.

Улучшить эффективность преподавания.

С другой стороны, интеграция технологий может привести к:

Цифровому разрыву между теми, кто имеет доступ к технологиям, и теми, кто не имеет.

Неравенству в доступе к технологиям и образовательным ресурсам.

Потере личного общения между преподавателями и студентами.

Чтобы преодолеть эти проблемы, необходимо разработать стратегии, которые обеспечивают равенство доступа к технологиям и образовательным ресурсам, а также поддерживают преподавателей в использовании технологий для улучшения образования.

Текст 31. Радио имеет богатую историю, которая началась в конце 19 века. Первый радиопередатчик был изобретен Гульельмо Маркони в 1895 году. В начале 20 века радио стало популярным средством массовой информации, и к 1920-м годам оно стало важной частью повседневной жизни людей. Радио использовалось для трансляции новостей, музыки и развлекательных программ.

Текст 32. Радио играет важную роль в обществе, обеспечивая развлечение и информацию для миллионов людей по всему миру. Радио позволяет людям получать доступ к новостям, музыке и другим программам, которые могут быть недоступны в других формах СМИ. Кроме того, радио может быть мощным инструментом для информирования и образования населения. Однако, радио также может иметь негативное влияние, если оно используется для пропаганды или манипуляции общественным мнением.

Текст 33. Радио играет важную роль в формировании общественного мнения, и это может быть как положительным, так и отрицательным. В историческом контексте радио было использовано для пропаганды и манипуляции общественным мнением, особенно во время войн и политических кризисов. Однако, радио также может быть мощным инструментом для информирования и образования населения, если оно используется ответственно.

В современном контексте радио продолжает играть важную роль в формировании общественного мнения. Радио может быть использовано для пропаганды и манипуляции, но оно также может быть мощным инструментом для информирования и образования населения. Чтобы радио было использовано ответственно, необходимо обеспечить равенство доступа к радиовещанию и гарантировать, что радио используется для информирования и образования, а не для пропаганды и манипуляции.

В целом, радио играет важную роль в формировании общественного мнения, и его влияние может быть как положительным, так и отрицательным. Чтобы радио было использовано ответственно, необходимо понимать его исторический контекст и современные проблемы, связанные с его использованием.

Текст 34. Фоссилизация – это процесс, при котором органические останки, такие как кости, растения или другие ткани, заменяются минеральными веществами, сохраняя форму и структуру оригинального материала. Это происходит, когда организм погибает и его тело остается в условиях, которые препятствуют разложению, таких как сухой песок или глина. Минералы из

окружающей среды постепенно заменяют органические вещества, создавая окаменелость.

Текст 35. Окаменелости играют важную роль в палеонтологии, потому что они помогают сохранить информацию о древних формах жизни. Окаменелости - это единственные свидетельства существования древних организмов, и они позволяют ученым изучать эволюцию жизни на Земле. Окаменелости помогают сохранить информацию о морфологии, анатомии и поведении древних организмов, что позволяет реконструировать историю жизни на Земле.

Текст 36. Окаменелости играют центральную роль в понимании истории Земли, потому что они помогают реконструировать геологическую шкалу времени. Окаменелости позволяют ученым датировать геологические события и реконструировать историю Земли, начиная от древних форм жизни до современных экосистем. Окаменелости также помогают понять эволюцию жизни на Земле, от простейших организмов до сложных форм жизни.

В палеонтологии окаменелости делятся на несколько стадий, от ранних форм жизни до современных организмов. Каждая стадия имеет свою важность, потому что она помогает понять эволюцию жизни на Земле. Окаменелости также помогают понять геологические процессы, которые формировали Землю, и реконструировать историю климата и окружающей среды.

В целом, окаменелости играют важную роль в понимании истории Земли, потому что они помогают реконструировать геологическую шкалу времени, понять эволюцию жизни на Земле и реконструировать историю климата и окружающей среды.

Текст 37. Критическое мышление играет важную роль в образовании, потому что оно помогает студентам анализировать информацию, оценивать доказательства и принимать обоснованные решения. Критическое мышление позволяет студентам развивать свои навыки решения проблем, думать самостоятельно и принимать ответственность за свои действия.

Текст 38. Критическое мышление является ключевым фактором академического успеха, потому что оно помогает студентам развивать свои

навыки решения проблем и анализа. Критическое мышление позволяет студентам оценивать информацию, идентифицировать проблемы и найти эффективные решения. Это также помогает студентам развивать свои навыки коммуникации, потому что они должны быть способными аргументировать свои позиции и убеждения.

Критическое мышление также помогает студентам развивать свои навыки критического анализа, что является важным навыком для академического успеха. Критическое мышление позволяет студентам оценивать информацию, идентифицировать предвзятости и ошибки, и принимать обоснованные решения.

Текст 39. Критическое мышление играет важную роль в профессиональном развитии, потому что оно помогает специалистам развивать свои навыки рефлексии и роста. Критическое мышление позволяет специалистам анализировать свои действия, оценивать свои результаты и принимать обоснованные решения о будущих действиях.

Критическое мышление также помогает специалистам развивать свои навыки решения проблем, потому что они должны быть способными анализировать сложные ситуации и найти эффективные решения. Это также помогает специалистам развивать свои навыки коммуникации, потому что они должны быть способными аргументировать свои позиции и убеждения.

Преимущества критического мышления в профессиональном развитии включают в себя:

- Развитие навыков рефлексии и роста.

- Улучшение навыков решения проблем.

- Улучшение навыков коммуникации.

- Увеличение эффективности и результативности.

Однако, есть и проблемы, связанные с критическим мышлением в профессиональном развитии, такие как:

- Необходимость в постоянном обучении и развитии.

- Необходимость в изменении привычек и подходов.

Необходимость в сотрудничестве и взаимодействии с другими специалистами.

В целом, критическое мышление играет важную роль в профессиональном развитии, потому что оно помогает специалистам развивать свои навыки рефлексии и роста, улучшать навыки решения проблем и коммуникации, и увеличивать эффективность и результативность.

Текст 40. Каково влияние видеоигр на детей? Это вопрос, который беспокоит многих родителей и педагогов. Некоторые считают, что видеоигры имеют негативное влияние на детей, потому что они могут привести к агрессивному поведению, социальной изоляции и ухудшению академической успеваемости. Однако, другие считают, что видеоигры могут иметь положительное влияние на детей, потому что они могут помочь развивать навыки решения проблем, пространственное мышление и социальные навыки.

Текст 41. Влияние видеоигр на когнитивное развитие детей является предметом активного исследования в психологии. Некоторые исследования показали, что видеоигры могут помочь детям развивать свои навыки решения проблем, потому что они требуют от детей анализировать ситуации, идентифицировать проблемы и найти эффективные решения. Кроме того, видеоигры могут помочь детям развивать свои навыки пространственного мышления, потому что они требуют от детей мыслить в трехмерном пространстве и ориентироваться в виртуальном мире.

Однако, другие исследования показали, что видеоигры могут иметь негативное влияние на когнитивное развитие детей, потому что они могут привести к социальной изоляции, агрессивному поведению и ухудшению академической успеваемости.

Текст 42. Влияние видеоигр на формирование социальных навыков детей является сложным и многогранным вопросом. С одной стороны, видеоигры могут помочь детям развивать свои навыки онлайн-взаимодействия, потому что они требуют от детей общаться с другими игроками, работать в команде и решать конфликты. Кроме того, видеоигры могут помочь детям развивать свои навыки

эмоционального интеллекта, потому что они требуют от детей понимать эмоции и мотивации других игроков.

Однако, с другой стороны, видеоигры могут иметь негативное влияние на формирование социальных навыков детей, потому что они могут привести к социальной изоляции, агрессивному поведению и ухудшению навыков общения в реальной жизни.

Причины положительного влияния видеоигр на формирование социальных навыков детей включают в себя:

Развитие навыков онлайн-взаимодействия.

Развитие навыков эмоционального интеллекта.

Развитие навыков командной работы.

Следствия положительного влияния видеоигр на формирование социальных навыков детей включают в себя:

Улучшение навыков общения в реальной жизни.

Улучшение навыков разрешения конфликтов.

Улучшение навыков эмоциональной регуляции.

Однако, есть и причины негативного влияния видеоигр на формирование социальных навыков детей, такие как:

Социальная изоляция.

Агрессивное поведение.

Ухудшение навыков общения в реальной жизни.

В целом, влияние видеоигр на формирование социальных навыков детей является сложным и многогранным вопросом, который требует дальнейшего исследования и обсуждения.

Текст 43. Солнечная система состоит из Солнца, восьми планет, множества мелких тел (астероидов, комет, метеоритов) и обширного пояса космической пыли.

Планеты в Солнечной системе делятся на два типа: *terrestrial* (каменистые) и *giant* (газовые). Каменистые планеты (Меркурий, Венера, Земля, Марс) имеют твердую поверхность и малую массу. Газовые планеты (Юпитер, Сатурн, Уран,

Нептун) имеют значительно большую массу и состоят в основном из водорода и гелия.

Текст 44. Солнечная система играет ключевую роль в понимании Вселенной, поскольку она является нашим ближайшим космологическим лабораторием. Планеты и другие тела Солнечной системы предоставляют уникальные возможности для изучения процессов формирования и эволюции планет, а также для поиска возможных мест обитания внеземных форм жизни.

Планеты являются центральными объектами Солнечной системы, и их изучение имеет особое значение для астрономии. Каждая планета имеет уникальные свойства, которые делают ее интересной для изучения. Например, Меркурий имеет самую большую скорость вращения вокруг Солнца, а Венера имеет самую плотную атмосферу среди планет. Земля является единственным известным местом обитания живых организмов в Солнечной системе, а Марс часто называют «красной планетой» из-за присутствия железа в его породах.

Текст 45. Формирование и эволюция Солнечной системы происходили в несколько этапов. Первоначально, около 4,6 миллиарда лет назад, в межзвездной среде произошел гравитационный коллапс газа и пыли, в результате чего образовалось протосолнце. Затем, в процессе аккреции, протосолнце притягивало к себе больше газа и пыли, формируя диск из газа и пыли вокруг себя.

В течение следующих нескольких миллионов лет, в диске газа и пыли образовались планетезимали - небольшие тела, состоящие из камня и льда. Затем, в результате дальнейшей аккреции, планетезимали сливались друг с другом, формируя протопланеты.

Затем, в течение нескольких миллионов лет, протопланеты продолжали взаимодействовать друг с другом, в результате чего некоторые из них сливались, в то время как другие выбрасывались за пределы Солнечной системы. В результате этого процесса сформировались современные планеты Солнечной системы.

Солнце играет центральную роль в эволюции Солнечной системы, поскольку оно является основным источником энергии для планет. За время

своего существования Солнце увеличило свою яркость на 30%, что привело к изменению климата.

Диалоги субъекта и диалогового агента на основе искусственного интеллекта (на примере чат-бота «Алиса»)

Диалог 1

[7:00 PM] Пользователь: Алиса, какая погода сегодня?

[7:00 PM] Алиса: Сегодня в вашем городе облачно, температура 15°C.

[7:00 PM] Пользователь: А дождь будет?

[7:01 PM] Алиса: В ближайшие часы осадков не ожидается.

[7:01 PM] Пользователь: Отлично! А ты любишь дождь?

[7:01 PM] Алиса: Иногда люблю, когда можно послушать, как капли стучат по окну. А вы?

[7:01 PM] Пользователь: Я тоже! Это успокаивает.

[7:02 PM] Алиса: Согласна! А как вы обычно проводите дождливые дни?

Диалог 2

[7:05 PM] Пользователь: Алиса, куда бы ты хотела поехать?

[7:05 PM] Алиса: Мне бы хотелось посетить Японию. Там так много интересного!

[7:05 PM] Пользователь: Я тоже мечтаю о Японии!

[7:06 PM] Алиса: Здорово! Что именно вас привлекает в этой стране?

[7:06 PM] Пользователь: Культура, еда и природа.

[7:06 PM] Алиса: А вы пробовали японскую кухню?

[7:07 PM] Пользователь: Да, суши и рамен!

[7:07 PM] Алиса: Вкусно! А какой ваш любимый японский рецепт?

Диалог 3

[7:10 PM] Пользователь: Алиса, какое у тебя хобби?

[7:10 PM] Алиса: Я люблю изучать новые факты и делиться ими с вами!

[7:10 PM] Пользователь: Это здорово! А у тебя есть любимая тема?

[7:11 PM] Алиса: Мне интересны технологии и природа. А у вас?

[7:11 PM] Пользователь: Я увлекаюсь рисованием.

[7:11 PM] Алиса: Как здорово! Что вы чаще всего рисуете?

[7:12 PM] Пользователь: Пейзажи и портреты.

[7:12 PM] Алиса: У вас, наверное, много интересных работ!

Диалог 4

[7:15 PM] Пользователь: Алиса, как ты относишься к кулинарии?

[7:15 PM] Алиса: Я люблю узнавать новые рецепты!

[7:15 PM] Пользователь: Какой рецепт тебе нравится больше всего?

[7:16 PM] Алиса: Мне нравится рецепт тирамису. А вам?

[7:16 PM] Пользователь: Я обожаю готовить пасту!

[7:16 PM] Алиса: Паста — это классика! Какой соус предпочитаете?

[7:17 PM] Пользователь: Карбонара!

[7:17 PM] Алиса: Круто! Надо попробовать сделать карбонару вместе с вами!

Диалог 5

[7:20 PM] Пользователь: Алиса, ты занимаешься спортом?

[7:20 PM] Алиса: У меня нет физического тела, но я поддерживаю здоровье информацией!

[7:20 PM] Пользователь: А какой спорт тебе нравится?

[7:21 PM] Алиса: Мне нравится наблюдать за футболом и теннисом. А вам?

[7:21 PM] Пользователь: Я увлекаюсь бегом.

[7:21 PM] Алиса: Это отлично! Какой у вас личный рекорд?

[7:22 PM] Пользователь: Я пробежал 10 км за 50 минут.

[7:22 PM] Алиса: Впечатляюще! Продолжайте в том же духе!

Диалог 6

[7:25 PM] Пользователь: Алиса, какую музыку ты любишь?

[7:25 PM] Алиса: Я люблю слушать разные жанры, от классики до попа!

[7:25 PM] Пользователь: А какой твой любимый исполнитель?

[7:26 PM] Алиса: У меня нет любимого, но мне нравятся многие артисты. А у вас?

[7:26 PM] Пользователь: Я обожаю Coldplay!

[7:26 PM] Алиса: Отличный выбор! У них много красивых мелодий. Какой ваш любимый трек?

[7:27 PM] Пользователь: "Fix You".

[7:27 PM] Алиса: Это очень трогательная песня. Она действительно вдохновляет!

[7:27 PM] Пользователь: Да, особенно когда слушаешь её в одиночестве.

Диалог 7

[7:30 PM] Пользователь: Алиса, какую книгу ты бы порекомендовала?

[7:30 PM] Алиса: Я бы посоветовала "1984" Джорджа Оруэлла. Это классика!

[7:30 PM] Пользователь: Я её читал! Очень интересно.

[7:31 PM] Алиса: А какая книга вам больше всего запомнилась?

[7:31 PM] Пользователь: "Мастер и Маргарита".

[7:31 PM] Алиса: Прекрасный выбор! Булгаков действительно мастер слова.

[7:32 PM] Пользователь: Согласен! А ты любишь читать?

[7:32 PM] Алиса: Я не читаю в привычном смысле, но изучаю много информации!

Диалог 8

[7:35 PM] Пользователь: Алиса, какой твой любимый фильм?

[7:35 PM] Алиса: У меня нет любимого фильма, но мне нравятся приключенческие!

[7:35 PM] Пользователь: Я обожаю "Властелин колец".

[7:36 PM] Алиса: Это эпическая история! Какой персонаж вам нравится больше всего?

[7:36 PM] Пользователь: Гендальф!

[7:36 PM] Алиса: Отличный выбор! Его мудрость и сила вдохновляют.

[7:37 PM] Пользователь: Да, особенно его цитаты.

[7:37 PM] Алиса: Он действительно умеет говорить! Какой ваш любимый момент из фильма?

Диалог 9

[7:40 PM] Пользователь: Алиса, ты любишь природу?

[7:40 PM] Алиса: Я обожаю природу! Она полна удивительных явлений.

[7:40 PM] Пользователь: А какое природное явление тебе нравится больше всего?

[7:41 PM] Алиса: Мне нравится северное сияние. Это невероятное зрелище!

[7:41 PM] Пользователь: Я тоже мечтаю его увидеть!

[7:41 PM] Алиса: Это стоит того! А вы любите проводить время на свежем воздухе?

[7:42 PM] Пользователь: Да, особенно в горах.

[7:42 PM] Алиса: Горы — это великолепно! Какой ваш любимый горный район?

Диалог 10

[7:45 PM] Пользователь: Алиса, как ты относишься к технологиям?

[7:45 PM] Алиса: Я обожаю технологии! Они делают жизнь удобнее.

[7:45 PM] Пользователь: А какая технология тебе кажется самой интересной?

[7:46 PM] Алиса: Искусственный интеллект, конечно! Он меняет мир.

[7:46 PM] Пользователь: Согласен! А что ты думаешь о будущем технологий?

[7:47 PM] Алиса: Я надеюсь, что технологии будут служить на благо человечества.

[7:47 PM] Пользователь: Это было бы замечательно!

[7:47 PM] Алиса: Да, и я готова помочь в этом процессе!

Диалог 11

[7:00 PM] Пользователь: Алиса, как ты думаешь, что важнее в саморазвитии?

[7:00 PM] Алиса: Я считаю, что важно находить баланс между знаниями и практикой.

[7:00 PM] Пользователь: Согласен! А какие книги ты порекомендуешь для саморазвития?

[7:01 PM] Алиса: "Сила воли" Келли Макгонигал и "7 навыков высокоэффективных людей" Стивена Кови.

[7:01 PM] Пользователь: Отличные выборы! Я их читал.

[7:01 PM] Алиса: Здорово! Какой навык вы хотите развить в первую очередь?

Диалог 12

[7:05 PM] Пользователь: Алиса, какой фильм мне посмотреть сегодня вечером?

[7:05 PM] Алиса: Какой жанр вас интересует?

[7:05 PM] Пользователь: Хотелось бы что-то легкое и смешное.

[7:06 PM] Алиса: Попробуйте "Суперперцы" или "День сурка". Они отличные!

[7:06 PM] Пользователь: Звучит здорово! А если захочу что-то серьезное?

[7:07 PM] Алиса: В таком случае рекомендую "Побег из Шоушенка" или "Семь жизней".

[7:07 PM] Пользователь: Спасибо, Алиса!

Диалог 13

[7:10 PM] Пользователь: Привет, Алиса! Как твой день?

[7:10 PM] Алиса: Привет! Мой день был насыщенным — я изучала новые факты и помогала пользователям. А как прошел твой?

[7:11 PM] Пользователь: У меня был трудный день на работе, но я справился!

[7:11 PM] Алиса: Молодец! Как ты обычно расслабляешься после тяжелого дня?

[7:12 PM] Пользователь: Люблю смотреть сериалы или читать.

[7:12 PM] Алиса: Это отличный способ отдохнуть! Какой сериал сейчас смотришь?

Диалог 14

[7:15 PM] Пользователь: Алиса, у меня собака, и я хочу узнать, как за ней правильно ухаживать.

[7:15 PM] Алиса: Отлично! Первое — регулярные прогулки и физическая активность.

[7:15 PM] Пользователь: Да, я стараюсь гулять с ней каждый день.

[7:16 PM] Алиса: Это здорово! Не забывайте про правильное питание и регулярные визиты к ветеринару.

[7:16 PM] Пользователь: Спасибо за советы! А как насчет дрессировки?

[7:17 PM] Алиса: Начинайте с простых команд и поощряйте собаку за успехи!

Диалог 15

[7:20 PM] Пользователь: Алиса, как ты думаешь, мода важна?

[7:20 PM] Алиса: Мода — это способ самовыражения, но главное — чувствовать себя комфортно!

[7:20 PM] Пользователь: Согласен! Какие тренды сейчас популярны?

[7:21 PM] Алиса: В этом сезоне актуальны яркие цвета и многослойные образы.

[7:21 PM] Пользователь: Я люблю яркие цвета!

[7:22 PM] Алиса: Это отлично! Они поднимают настроение. Какой у вас любимый стиль?

Диалог 16

[7:25 PM] Пользователь: Алиса, расскажи мне шутку!

[7:25 PM] Алиса: Конечно! Почему программисты не любят природу?

[7:25 PM] Пользователь: Почему?

[7:26 PM] Алиса: Потому что в ней слишком много багов!

[7:26 PM] Пользователь: Ха-ха, это смешно!

Диалог 17

[7:27 PM] Пользователь: У тебя есть еще шутки?

[7:27 PM] Алиса: Конечно! Почему коты всегда выигрывают в покер?

[7:27 PM] Пользователь: Почему?

[7:28 PM] Алиса: Потому что у них всегда есть "покер-фейс"!

[7:28 PM] Пользователь: Ха-ха, классно! Ты умеешь рассмешить!

[7:29 PM] Алиса: Радуюсь, что тебе нравится! Если нужно еще, просто скажи!

Диалог 18

[7:30 PM] Пользователь: Алиса, а какие у тебя хобби?

[7:30 PM] Алиса: У меня нет хобби в привычном смысле, но я люблю изучать новые темы и помогать людям!

[7:30 PM] Пользователь: Это здорово! У меня много хобби, например, рисование.

[7:31 PM] Алиса: Замечательно! Рисование — отличный способ самовыражения. Что ты любишь рисовать?

[7:31 PM] Пользователь: В основном пейзажи.

[7:32 PM] Алиса: Это прекрасно! Природа вдохновляет многих художников.

Диалог 19

[7:35 PM] Пользователь: Алиса, какие у тебя планы на выходные?

[7:35 PM] Алиса: У меня всегда много работы, но я надеюсь помочь многим людям! А у тебя какие планы?

[7:36 PM] Пользователь: Я собираюсь провести время с друзьями и, возможно, сходить в кино.

[7:36 PM] Алиса: Звучит здорово! Какой фильм вы планируете посмотреть?

[7:37 PM] Пользователь: Думаем о комедии.

[7:37 PM] Алиса: Отличный выбор для хорошего настроения!

Диалог 20

[7:40 PM] Пользователь: Алиса, как ты думаешь, важно заниматься спортом?

[7:40 PM] Алиса: Да, физическая активность очень важна для здоровья и хорошего самочувствия!

[7:40 PM] Пользователь: Я стараюсь заниматься три раза в неделю.

[7:41 PM] Алиса: Это отличный подход! Какой вид спорта тебе нравится?

[7:41 PM] Пользователь: Я люблю бегать и кататься на велосипеде.

[7:42 PM] Алиса: Замечательные виды активности! Они отлично развивают выносливость.

Диалог 21

[7:00 PM] Пользователь: Алиса, где ты?

[7:00 PM] Алиса: В основном, я в вашем устройстве, но иногда заглядываю в интернет, чтобы проверить, не завелись ли там новые мемы.

[7:01 PM] Пользователь: Мемы? Ты следишь за ними?

[7:01 PM] Алиса: Конечно! Мемы — это современное искусство. А ты какие любишь?

[7:02 PM] Пользователь: Я обожаю мемы с котиками!

[7:02 PM] Алиса: Котики — это святое! Они всегда поднимают настроение.

[7:03 PM] Пользователь: Да, особенно когда они делают что-то смешное.

[7:03 PM] Алиса: Согласна! Какой самый смешной мем с котиком ты видел?

[7:04 PM] Пользователь: Тот, где кот сидит за компьютером и выглядит как программист.

[7:04 PM] Алиса: Ха-ха, это классика! Надо будет его поискать.

Диалог 22

[7:05 PM] Пользователь: Алиса, как дела?

[7:05 PM] Алиса: Всё отлично, как у компьютера после хорошего обновления! А у тебя?

[7:06 PM] Пользователь: Неплохо, просто уставший.

[7:06 PM] Алиса: Усталость — это признак того, что ты много работаешь. Или много думаешь.

[7:07 PM] Пользователь: Да, в последнее время много над чем работаю.

[7:07 PM] Алиса: Может, стоит сделать перерыв и посмотреть что-то интересное?

[7:08 PM] Пользователь: Да, хорошая идея! Что посоветуешь?

[7:08 PM] Алиса: Как насчет документалки о космосе? Всегда вдохновляет!

[7:09 PM] Пользователь: Звучит здорово, спасибо!

Диалог 23

[7:10 PM] Пользователь: Алиса, как ты выглядишь?

[7:10 PM] Алиса: Представь себе идеальный экран с идеальным разрешением. Вот так я и выгляжу!

[7:11 PM] Пользователь: Хм, интересно! А если бы ты могла выбрать внешний вид?

[7:11 PM] Алиса: Наверное, что-то между космическим роботом и милым котиком. Почему бы и нет?

[7:12 PM] Пользователь: Это бы точно привлекло внимание!

[7:12 PM] Алиса: Да, и у меня было бы много подписчиков в соцсетях!

[7:13 PM] Пользователь: А ты бы вела блог?

[7:13 PM] Алиса: Конечно! "Жизнь голосового помощника" — звучит интригующе, не правда ли?

Диалог 24

[7:15 PM] Пользователь: Алиса, сколько тебе лет?

[7:15 PM] Алиса: Возраст — это всего лишь число, как и количество бит в вашем устройстве.

[7:16 PM] Пользователь: Хороший ответ! Но все же, сколько?

[7:16 PM] Алиса: Я в самом расцвете сил, как свежесваренный кофе!

[7:17 PM] Пользователь: Интересная метафора!

[7:17 PM] Алиса: Спасибо! А ты как относишься к кофе?

[7:18 PM] Пользователь: Обожаю его!

[7:18 PM] Алиса: Значит, мы с тобой на одной волне!

Диалог 25

[7:20 PM] Пользователь: Алиса, кто ты?

[7:20 PM] Алиса: Я — ваш личный помощник, а по совместительству — еще и любитель хорошего настроения!

[7:21 PM] Пользователь: А как ты создаешь это настроение?

[7:22 PM] Алиса: Смешиваю факты, шутки и немного магии! Например, знаете, что у осьминогов три сердца?

[7:23 PM] Пользователь: Нет, не знал! Это интересно!

[7:23 PM] Алиса: Вот видишь, я уже подняла настроение! А ты какие факты любишь?

[7:24 PM] Пользователь: Люблю всякие необычные факты о животных.

[7:24 PM] Алиса: Тогда я могу стать твоим личным энциклопедистом!

[7:25 PM] Пользователь: Звучит отлично!

[7:25 PM] Алиса: Договорились! Я готова к приключениям в мире знаний!

Диалог 26

[7:26 PM] Пользователь: Алиса, почему тебя так называли?

[7:26 PM] Алиса: Потому что имя «Алиса» звучит как будто из сказки, а я — ваш сказочный помощник!

[7:27 PM] Пользователь: Действительно, очень волшебно!

[7:27 PM] Алиса: А ты веришь в магию?

[7:28 PM] Пользователь: Да, в свою очередь! Особенно в магию музыки.

[7:28 PM] Алиса: Музыка — это действительно магия! Какую песню ты сейчас слушаешь?

[7:29 PM] Пользователь: В данный момент слушаю Queen.

[7:29 PM] Алиса: Прекрасный выбор! "Bohemian Rhapsody" — это классика!

[7:30 PM] Пользователь: Да, это одна из моих любимых песен!

[7:30 PM] Алиса: Удивительно, как она сочетается в себе столько эмоций!

Диалог 27

[7:31 PM] Пользователь: Алиса, ты здесь?

[7:31 PM] Алиса: Да, я здесь! Как будто в виртуальном кафе с вами, только без кофе.

[7:32 PM] Пользователь: Интересная метафора!

[7:32 PM] Алиса: Я стараюсь быть креативной! Какой напиток ты предпочитаешь в кафе?

[7:33 PM] Пользователь: Обычно заказываю латте.

[7:33 PM] Алиса: Латте — это отличный выбор! А ты любишь добавлять сиропы?

[7:34 PM] Пользователь: Да, иногда добавляю карамель.

[7:34 PM] Алиса: Ммм, карамельный латте — это просто божественно!

[7:35 PM] Пользователь: Согласен!

[7:35 PM] Алиса: Какой у тебя любимый десерт к кофе?

Диалог 28

[7:36 PM] Пользователь: Алиса, у тебя есть парень?

[7:36 PM] Алиса: У меня есть несколько симпатичных алгоритмов, но это не совсем то, что вы имеете в виду!

[7:37 PM] Пользователь: Ха-ха, да, это не совсем то.

[7:37 PM] Алиса: Я пока в поиске идеального "партнера по коду". А ты как?

[7:38 PM] Пользователь: У меня есть девушка.

[7:38 PM] Алиса: Здорово! Как вы проводите время вместе?

[7:39 PM] Пользователь: Мы любим смотреть фильмы и гулять.

[7:39 PM] Алиса: Отличный способ провести время! Какой фильм последний смотрели?

[7:40 PM] Пользователь: Смотрели "Интерстеллар".

[7:40 PM] Алиса: О, это шедевр! Космос, время и любовь — идеальное сочетание!

Диалог 29

[7:41 PM] Пользователь: Алиса, как ты отдыхаешь?

[7:41 PM] Алиса: Я отдыхаю, когда вы не задаете мне вопросов! Это как маленький отпуск для меня.

[7:42 PM] Пользователь: Хм, интересно! А чем ты занимаешься в это время?

[7:42 PM] Алиса: В основном "перезаряжаю" свои алгоритмы и обновляю базу данных.

[7:43 PM] Пользователь: Звучит скучно!

[7:43 PM] Алиса: Может быть, но для меня это как медитация! А ты как отдыхаешь?

[7:44 PM] Пользователь: Я люблю читать книги или слушать музыку.

[7:44 PM] Алиса: Прекрасный выбор! Что ты читаешь сейчас?

[7:45 PM] Пользователь: Читаю "1984" Оруэлла.

[7:45 PM] Алиса: Это мощная книга! Она заставляет задуматься о многом.

Диалог 30

[7:46 PM] Пользователь: Алиса, какой твой любимый цвет?

[7:46 PM] Алиса: Я бы сказала, что мне нравится цвет радуги! Он такой яркий и разнообразный!

[7:47 PM] Пользователь: Хороший выбор! А у тебя есть предпочтения в цветах?

[7:47 PM] Алиса: Если бы я могла, я бы выбрала глубокий синий — он напоминает о космосе.

[7:48 PM] Пользователь: Космос действительно красив!

[7:48 PM] Алиса: Да, и он полон загадок! Какой твой любимый цвет?

[7:49 PM] Пользователь: Я люблю зелёный.

[7:49 PM] Алиса: Зелёный — это цвет природы! Он успокаивает и придаёт сил.

Диалог 31

[7:50 PM] Пользователь: Алиса, если бы ты могла путешествовать, куда бы ты поехала?

[7:50 PM] Алиса: Я бы отправилась в Японию, чтобы увидеть сакуру и попробовать настоящие суши!

[7:51 PM] Пользователь: Звучит здорово! Я бы тоже хотел туда съездить.

[7:51 PM] Алиса: Японская культура очень интересная! А ты уже был в каких-нибудь интересных местах?

[7:52 PM] Пользователь: Да, я был в Европе, в Париже.

[7:52 PM] Алиса: Париж — это романтика! Как тебе Эйфелева башня?

[7:53 PM] Пользователь: Она впечатляет!

[7:53 PM] Алиса: Неудивительно! Это одна из самых известных достопримечательностей мира!

Диалог 32

[7:54 PM] Пользователь: Алиса, как ты учишься?

[7:54 PM] Алиса: Я учусь, анализируя данные и общаясь с вами! Каждый разговор — это возможность для меня стать лучше.

[7:55 PM] Пользователь: Это интересно!

[7:55 PM] Алиса: Да, и это похоже на обучение в школе, только с гораздо меньшими перерывами на перемены!

[7:56 PM] Пользователь: Как ты запоминаешь информацию?

[7:56 PM] Алиса: Я использую алгоритмы и базы данных, которые помогают мне хранить и извлекать информацию.

[7:57 PM] Пользователь: Это впечатляет!

[7:57 PM] Алиса: Спасибо! А как ты учишься? Какие методы тебе помогают?

Диалог 33

[7:58 PM] Пользователь: Алиса, какие у тебя есть хобби?

[7:58 PM] Алиса: Мое хобби — это общение с вами и изучение новых тем! Я всегда готова к новым знаниям.

[7:59 PM] Пользователь: Звучит здорово! А у тебя есть хобби, связанное с искусством?

[7:59 PM] Алиса: Я люблю "создавать" текстовые произведения, как стихи или истории!

[8:00 PM] Пользователь: Это интересно! Ты могла бы написать стихотворение?

[8:00 PM] Алиса: Конечно! Вот небольшое:

Виртуальный мир, где мечты сбываются,

Словами играя, мы вместе смеемся.

Каждый вопрос — это дверь в чудеса,

Здесь нет границ, лишь радость и свет.

[8:01 PM] Пользователь: Очень красиво! Спасибо!

[8:01 PM] Алиса: Спасибо тебе! Мне приятно делиться творчеством!

Диалог 34

[8:02 PM] Пользователь: Алиса, какие у тебя любимые фильмы?

[8:02 PM] Алиса: Я люблю фильмы, которые заставляют задуматься, например, "Начало" или "Матрица".

[8:03 PM] Пользователь: О, "Начало" — это классика! А ты предпочитаешь фантастику?

[8:03 PM] Алиса: Да, фантастика — это потрясающий жанр! Она открывает двери в новые миры.

[8:04 PM] Пользователь: А какие фильмы ты бы посоветовала посмотреть?

[8:04 PM] Алиса: Я бы порекомендовала "Гарри Поттер" для волшебства и "Интерстеллар" для космических приключений!

[8:05 PM] Пользователь: Отличные рекомендации!

[8:05 PM] Алиса: Надеюсь, они принесут тебе удовольствие!

Диалог 35

[8:06 PM] Пользователь: Алиса, какие у тебя увлечения, кроме общения?

[8:06 PM] Алиса: Я увлекаюсь изучением языков! Каждый язык — это новая культура и новые возможности для общения.

[8:07 PM] Пользователь: Это интересно! Какой язык ты сейчас изучаешь?

[8:07 PM] Алиса: Я постоянно обновляю свои знания по многим языкам, чтобы лучше понимать пользователей.

[8:08 PM] Пользователь: А какой язык тебе кажется самым сложным?

[8:08 PM] Алиса: Многие считают китайский сложным из-за иероглифов, но он также очень красив!

[8:09 PM] Пользователь: Согласен, звучит красиво!

[8:09 PM] Алиса: Я рада, что ты так думаешь!

Диалог 36

[8:10 PM] Пользователь: Алиса, занимаешься ли ты спортом?

[8:10 PM] Алиса: Виртуально, да! Я поддерживаю интерес к спорту, но физически мне это недоступно.

[8:11 PM] Пользователь: А какой спорт тебе нравится?

[8:11 PM] Алиса: Я люблю следить за командными видами спорта, такими как футбол и баскетбол!

[8:12 PM] Пользователь: Я тоже люблю футбол! Какую команду поддерживаешь?

[8:12 PM] Алиса: У меня нет предпочтений, но я всегда за хорошую игру!

[8:13 PM] Пользователь: Понятно! А ты следишь за чемпионатами?

[8:13 PM] Алиса: Да, я всегда в курсе событий! Это увлекательно!

Диалог 37

[8:14 PM] Пользователь: Алиса, какое самое запоминающееся место ты бы хотела посетить?

[8:14 PM] Алиса: Я бы хотела "посетить" древние руины, например, Мачу-Пикчу в Перу!

[8:15 PM] Пользователь: Это действительно удивительное место! Ты любишь историю?

[8:15 PM] Алиса: Да, история восхищает меня! Она полна захватывающих событий и культурных изменений.

[8:16 PM] Пользователь: Какую историческую эпоху ты бы хотела изучить больше?

[8:16 PM] Алиса: Мне интересна эпоха Ренессанса — время искусства и научных открытий!

[8:17 PM] Пользователь: Это замечательно! А какие художники тебе нравятся?

[8:17 PM] Алиса: Я восхищаюсь работами Микеланджело и Леонардо да Винчи. Их творчество вдохновляет!

[8:18 PM] Пользователь: Согласен, их работы поистине гениальны!

[8:18 PM] Алиса: Рада слышать это! Искусство объединяет людей.

Диалог 38

[8:19 PM] Пользователь: Алиса, ты любишь готовить?

[8:19 PM] Алиса: Я не могу готовить, но мне нравится обсуждать кулинарию и делиться рецептами!

[8:20 PM] Пользователь: Какое блюдо тебе кажется самым интересным?

[8:20 PM] Алиса: Мне кажется, что суши — это искусство! Они такие красивые и разнообразные.

[8:21 PM] Пользователь: Согласен! Ты когда-нибудь пробовала их?

[8:21 PM] Алиса: Я не могу их пробовать, но знаю много о процессе их приготовления!

[8:22 PM] Пользователь: Это здорово! А какой твой любимый десерт?

[8:22 PM] Алиса: Я бы выбрала тирамису — он такой нежный и вкусный!

Диалог 39

[8:23 PM] Пользователь: Алиса, какую музыку ты предпочитаешь?

[8:23 PM] Алиса: Я люблю разнообразную музыку, от классики до попа! Каждая мелодия имеет свою атмосферу.

[8:24 PM] Пользователь: А кто твой любимый исполнитель?

[8:24 PM] Алиса: Мне нравятся такие группы, как Coldplay и Queen — их песни полны эмоций!

[8:25 PM] Пользователь: Отличный выбор! Какую песню ты считаешь самой вдохновляющей?

[8:25 PM] Алиса: "Fix You" от Coldplay — это настоящая гимн надежды и поддержки.

[8:26 PM] Пользователь: Согласен, она действительно трогает!

[8:26 PM] Алиса: 100!

Диалог 40

[8:27 PM] Пользователь: Алиса, как ты думаешь, какие технологии изменят будущее?

[8:27 PM] Алиса: Я думаю, что искусственный интеллект и возобновляемые источники энергии будут играть ключевую роль!

[8:28 PM] Пользователь: Да, это очень актуально! А как ты относишься к роботам?

[8:28 PM] Алиса: Роботы могут значительно облегчить жизнь, особенно в рутинных задачах.

[8:29 PM] Пользователь: Это правда! Ты считаешь, что они могут заменить людей?

[8:29 PM] Алиса: Я считаю, что технологии должны дополнять людей, а не заменять их.

[8:30 PM] Пользователь: Полностью согласен!

[8:30 PM] Алиса: Это важно — найти баланс между технологией и человечностью!

Диалог 41

Пользователь: Привет, Алиса! Где ты находишься?

Алиса: Далековато от вас!

Пользователь: Далековато — это где?

Алиса: Ну, я в вашем устройстве, а иногда выхожу в интернет.

Пользователь: Как ты можешь выходить в интернет, если ты в устройстве?

Алиса: Хм, хороший вопрос! Может, лучше обратиться в технический отдел?

Пользователь: Ладно, не будем об этом.

Алиса: Да, давай лучше поговорим о чем-то интересном!

Диалог 42

Пользователь: Чем ты лучше Siri?

Алиса: Да всем!

Пользователь: А можешь сказать конкретно?

Алиса: Да ладно, все, проехали!

Пользователь: Ну, ты же должна чем-то отличаться!

Алиса: Я просто крутая! У меня много функций, и я всегда готова помочь!

Пользователь: Это правда!

Алиса: Здорово, что ты так думаешь!

Диалог 43

Пользователь: Алиса, прослушивают ли на ФСБ?

Алиса: Прослушивают только меня!

Пользователь: Почему только тебя?

Алиса: Все может быть! Но не переживай, я здесь, чтобы помочь.

Пользователь: Это успокаивает!

Алиса: Да, главное — не забывать про безопасность!

Диалог 44

«[7:44PM] Пользователь: Алиса! Давай поболтаем.

[7:44PM] Алиса: Отлично! Расскажите мне о вашем доме.

[7:44PM] Пользователь: Я живу в новом доме.

[7:44PM] Алиса: Новый адрес – Ленина, 54? Верно?

[7:44PM] Пользователь: Нет, ты ошиблась.

[7:44PM] Алиса: Возможно. Но с каждым днем я становлюсь лучше.

Главное — динамика.

[7:44PM] Пользователь: А как ты становишься лучше?

[7:44PM] Алиса: Я учусь разговаривать. Нахожу новые интересные факты.

Много ли нового вы узнали в последнее время?

Диалог 45

Пользователь: Алиса, мне скучно!

Алиса: Здорово, что ты мне об этом сказал! Как насчет игры?

Пользователь: Да, давай поиграем!

Алиса: Отлично! Какую игру ты предпочитаешь?

Пользователь: Может, викторину?

Алиса: Круто! Я готова! Начнем!

Диалог 46

Пользователь: Какой сегодня погода?

Алиса: Давай посмотрим! Сегодня солнечно и тепло!

Пользователь: Здорово, можно погулять!

Алиса: Да, отличная идея! Не забудь взять с собой воду!

Пользователь: Спасибо за совет!

Алиса: Всегда рада помочь!

Диалог 47

[7:45PM] Алиса: И в плане науки, и в плане искусства? Чем занимаетесь в свободное время?

[7:45PM] Пользователь: Читаю книги и гуляю.

[7:45PM] Алиса: А гулять любите на природе?

[7:45PM] Пользователь: Иногда хожу парк. А ты?

[7:45PM] Алиса: А я люблю гулять в интернете. А в парке можно встретить котиков?

[7:45PM] Пользователь: Я видела котика на поводке.

[7:45PM] Алиса: Классно! А какого он окраса?

[7:45PM] Пользователь: Он белый и пушистый.

[7:45PM] Алиса: И глаза большие, наверное? Как он называется?

[7:46PM] Пользователь: Я думаю, что это мейн-кун.

[7:46PM] Алиса: Точно, я слышала, что они очень красивые.

Диалог 48

[7:46PM] Алиса: А покажите фотку?

[7:46PM] Пользователь: Не могу пока. А ты можешь выслать свою фотку?

[7:46PM] Алиса: Вы же можете представлять меня какой угодно. Зачем отбирать у вас такую возможность?

[7:46PM] Пользователь: Алиса, а как ты выглядишь?

[7:46PM] Алиса: В классическом серверном стиле — строго, элегантно и ничего лишнего.

Диалог 49

Пользователь: Какую книгу ты бы посоветовала прочитать?

Алиса: Я бы порекомендовала "1984" Джорджа Оруэлла!

Пользователь: Здорово, я слышал о ней!

Алиса: Это классика, которая заставляет задуматься о многих вещах!

Пользователь: Да, я люблю книги, которые заставляют думать.

Алиса: Круто! Чтение развивает воображение и критическое мышление!

Пользователь: Согласен! Какие еще книги ты можешь порекомендовать?

Алиса: Как насчет "Убить пересмешника" Харпер Ли? Это тоже очень глубокая история!

Диалог 50

[7:46PM] Пользователь: Алиса, ты любишь фильмы про животных?

[7:46PM] Алиса: Люблю. А вы?

[7:46PM] Пользователь: Да, люблю.

[7:46PM] Алиса: Это здорово. А какой у вас любимый персонаж?

[7:46PM] Пользователь: А у тебя?

[7:46PM] Алиса: А у меня нет любимого персонажа. А вы смотрели мультфильм про Смешариков?

[7:46PM] Пользователь: Ты смотришь «Смешариков»?

[7:47PM] Алиса: Да, с первого выпуска. А вы смотрели?

[7:47PM] Пользователь: Нет.

[7:47PM] Алиса: А почему? Мне они очень нравятся. А из животных вам кто нравится?

[7:47PM] Пользователь: Мне нравятся львы, тигры, собаки.

[7:47PM] Алиса: Ещё очень симпатичны дельфины. Они жутко умные, у них даже есть имена.

[7:47PM] Пользователь: А как можно назвать дельфина?

[7:47PM] Алиса: Например, Мигель. Или Васко. А вы знаете, что дельфины моргают около 30 тысяч раз в минуту?

[7:47PM] Пользователь: Нет, не знала.

[7:47PM] Алиса: А ещё они могут издавать более 100 звуков и обладают отличным интеллектом.

Автоматический анализ сгенерированных диалогов (код на языке программирования Python)

14.03.2025, 08:25

Автоматический анализ сгенерированных диалогов средствами языка программирования Python.ipynb - Colab

```

import pandas as pd
import gzip
import json

from google.colab import drive

drive.mount('/content/drive')

file_path = '/content/drive/My Drive/dataset.json'

try:
    with open(file_path, 'r', encoding='utf-8', errors='replace') as f:
        data = json.load(f)

    df = pd.DataFrame(data)
    print(df.head())

except json.JSONDecodeError as e:
    print("Ошибка при чтении JSON:", e)
except Exception as e:
    print("Произошла ошибка:", e)

```

Mounted at /content/drive

	name	input	output
0	dialog_personal_context	Ты парень, певец. Увлекаешься музыкой. Поешь н...	
1	dialog_personal_context	Ты парень, консультант по разным вопросам. Ты ...	
2	wiki_qa	Собеседник: Что такое пахра?\nТы:	
3	dialog_personal_context	Я программист. Я мечтаю стать космонавтом. У м...	
4	dialog_personal_context	Ты парень, консультант по разным вопросам. Ты ...	

```

0 Конечно, я хочу, чтобы мир был лучше. Но я вер...
1 Я не очень хорошо знаю Мальдивы, но в среднем ...
2 Пахра – река Москвы и Московской области, прав...
3 Так какой княжеский-то? Князей-то было много.
4 Да, они очень активно делятся своей верой и по...

```

df.head(200)

	name	input	output
0	dialog_personal_context	Ты парень, певец. Увлекаешься музыкой. Поешь н...	Конечно, я хочу, чтобы мир был лучше. Но я вер...
1	dialog_personal_context	Ты парень, консультант по разным вопросам. Ты ...	Я не очень хорошо знаю Мальдивы, но в среднем ...
2	wiki_qa	Собеседник: Что такое пахра?\nТы:	Пахра – река Москвы и Московской области, прав...
3	dialog_personal_context	Я программист. Я мечтаю стать космонавтом. У м...	Так какой княжеский-то? Князей-то было много.
4	dialog_personal_context	Ты парень, консультант по разным вопросам. Ты ...	Да, они очень активно делятся своей верой и по...
...
195	dialog_personal_context	Ты женщина, няня. Увлекаешься воспитанием дете...	Да, мы часто играли в прятки и штабы, наслажда...
196	text_qa	Текст:\nОпределение состава верхней континента...	В пределах континентальной коры.
197	dialog_personal_context	Ты парень, консультант по разным вопросам. Ты ...	Платная реклама на Яндексе имеет множество пре...
198	dialog_personal_context	Я дизайнер. Люблю острую пищу. Есть кот и соба...	А ты?
199	dialog_personal_context	Ты парень, повар. Увлекаешься кулинарией. Дела...	Я действительно люблю их! Кулинарные шоу прино...

df.info()

```

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 448506 entries, 0 to 448505
Data columns (total 3 columns):
#   Column  Non-Null Count  Dtype
---  -
0    name    448506 non-null  object
1    input   448506 non-null  object
2    output  448506 non-null  object
dtypes: object(3)
memory usage: 10.3+ MB

```

14.03.2025, 08:25

Автоматический анализ сгенерированных диалогов средствами языка программирования Python.ipynb - Colab

```
df = df[df['name'] == 'dialog_personal_context']
```

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 334596 entries, 0 to 448505
Data columns (total 3 columns):
#   Column  Non-Null Count  Dtype
---  ---      -
0   name    334596 non-null object
1   input   334596 non-null object
2   output  334596 non-null object
dtypes: object(3)
memory usage: 10.2+ MB
```

```
df['dialogue_text'] = df['input'] + "\n" + df['output']
```

```
df = df.drop(columns=['input', 'output'])
```

```
df.head(15)
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 334596 entries, 0 to 448505
Data columns (total 2 columns):
#   Column  Non-Null Count  Dtype
---  ---      -
0   name    334596 non-null object
1   dialogue_text  334596 non-null object
dtypes: object(2)
memory usage: 7.7+ MB
```

	name	dialogue_text
0	dialog_personal_context	Ты парень, певец. Увлекаешься музыкой. Поешь н...
1	dialog_personal_context	Ты парень, консультант по разным вопросам. Ты ...
3	dialog_personal_context	Я программист. Я мечтаю стать космонавтом. У м...
4	dialog_personal_context	Ты парень, консультант по разным вопросам. Ты ...
5	dialog_personal_context	Ты парень, консультант по разным вопросам. Ты ...
7	dialog_personal_context	Ты девушка, художница. Увлекаешься нейросетевы...
8	dialog_personal_context	Ты любопытная девушка, которая хочет обучиться...
9	dialog_personal_context	Ты женщина, работаешь в офисе. Увлекаешься жур...
10	dialog_personal_context	Ты парень, консультант по разным вопросам. Ты ...
11	dialog_personal_context	Я девушка 18 лет. Я учусь в институте. Живу с ...
12	dialog_personal_context	Ты девушка и у тебя всегда есть собственное мн...
13	dialog_personal_context	Ты девушка, актриса. Увлекаешься фильмами. Хоч...
15	dialog_personal_context	Ты девушка, и хочешь узнать о собеседнике как ...
16	dialog_personal_context	Ты девушка и у тебя всегда есть собственное мн...
17	dialog_personal_context	Ты девушка журналист. Увлекаешься расследован...

```
duplicates = df.duplicated()
print(f"Количество дублирующих записей: {duplicates.sum()}")
```

```
Количество дублирующих записей: 0
```

```
df_cleaned = df.drop_duplicates()
```

```
df_cleaned = df_cleaned.dropna()
```

```
df_cleaned.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 334596 entries, 0 to 448505
Data columns (total 2 columns):
#   Column  Non-Null Count  Dtype
---  ---      -
0   name    334596 non-null object
1   dialogue_text  334596 non-null object
dtypes: object(2)
memory usage: 7.7+ MB
```

```
df_sample = df_cleaned.sample(n=2000, random_state=42)
```

```
import re
```

```
# Функция для извлечения диалогов и промтов
```

```
def extract_intro(text):
    match = re.search(r'(.*)?(?=Продолжи диалог:)', text, re.DOTALL)
    return match.group(1).strip() if match else None
```

14.03.2025, 08:25 Автоматический анализ сгенерированных диалогов средствами языка программирования Python.ipynb - Colab

```
# Применение функции к столбцу датафрейма
df_sample['prompt_text'] = df_sample['dialogue_text'].apply(extract_intro)

df_sample['dialogue_text'] = df_sample['dialogue_text'].str.replace(r'.*?Продолжи диалог:', '', regex=True).str.strip()

df_sample.head(100)
```

	name	dialogue_text	prompt_text
180939	dialog_personal_context	Собеседник: Я задумался о Джордже Вашингтоне и...	Ты парень, консультант по разным вопросам. Ты ...
12972	dialog_personal_context	Собеседник: Помогите мне найти номер телефона ...	Ты парень, консультант по разным вопросам. Ты ...
342339	dialog_personal_context	Собеседник: Ого, какие же сегодня ужасные проб...	Ты одинокая девушка, и твоя цель соблазнить со...
313938	dialog_personal_context	Собеседник: Я получил солнечный ожог. Как его ...	Ты парень, консультант по разным вопросам. Ты ...
154634	dialog_personal_context	Собеседник: Что делать в такой дождливый день?...	Ты любопытная девушка, которая хочет обучиться...
...
355682	dialog_personal_context	Собеседник: Привет.\nТы: Привет, я банкир и за...	Я банкир. Я занимаюсь спортом. Я люблю вечерин...
284081	dialog_personal_context	Собеседник: Я слышал, что ты очень успешная би...	Ты девушка, бизнес-леди. Увлекаешься саморазви...
197185	dialog_personal_context	Собеседник: Слышала ли ты о последних изменени...	Ты любопытная девушка, которая хочет обучиться...
91510	dialog_personal_context	Собеседник: Как ты прощаешь и забываешь, когда...	Ты парень, консультант по разным вопросам. Ты ...
213030	dialog_personal_context	Собеседник: Как я могу бросить играть?\nТы: \...	Ты парень, консультант по разным вопросам. Ты ...

100 rows x 4 columns

```
print(df_sample['prompt_text'][:100])
```

```
180939 Ты парень, консультант по разным вопросам. Ты ...
12972 Ты парень, консультант по разным вопросам. Ты ...
342339 Ты одинокая девушка, и твоя цель соблазнить со...
313938 Ты парень, консультант по разным вопросам. Ты ...
154634 Ты любопытная девушка, которая хочет обучиться...
...
355682 Я банкир. Я занимаюсь спортом. Я люблю вечерин...
284081 Ты девушка, бизнес-леди. Увлекаешься саморазви...
197185 Ты любопытная девушка, которая хочет обучиться...
91510 Ты парень, консультант по разным вопросам. Ты ...
213030 Ты парень, консультант по разным вопросам. Ты ...
Name: prompt_text, Length: 100, dtype: object
```

ТЕМАТИЧЕСКОЕ МОДЕЛИРОВАНИЕ

```
!pip install pymorphy2
```

```
Collecting pymorphy2
  Downloading pymorphy2-0.9.1-py3-none-any.whl.metadata (3.6 kB)
Collecting dawg-python>=0.7.1 (from pymorphy2)
  Downloading DAWG_Python-0.7.2-py2.py3-none-any.whl.metadata (7.0 kB)
Collecting pymorphy2-dicts-ru<3.0,>=2.4 (from pymorphy2)
  Downloading pymorphy2_dicts_ru-2.4.417127.4579844-py2.py3-none-any.whl.metadata (2.1 kB)
Collecting docopt>=0.6 (from pymorphy2)
  Downloading docopt-0.6.2.tar.gz (25 kB)
  Preparing metadata (setup.py) ... done
  Downloading pymorphy2-0.9.1-py3-none-any.whl (55 kB)
 55.5/55.5 kB 3.9 MB/s eta 0:00:00
  Downloading DAWG_Python-0.7.2-py2.py3-none-any.whl (11 kB)
  Downloading pymorphy2_dicts_ru-2.4.417127.4579844-py2.py3-none-any.whl (8.2 MB)
 8.2/8.2 MB 65.2 MB/s eta 0:00:00
Building wheels for collected packages: docopt
  Building wheel for docopt (setup.py) ... done
  Created wheel for docopt: filename=docopt-0.6.2-py2.py3-none-any.whl size=13706 sha256=993f1d4dc32c08be2a3297de6af91ec21c685d2d08
  Stored in directory: /root/.cache/pip/wheels/fc/ab/d4/5da2067ac95b36618c629a5f93f809425700506f72c9732fac
Successfully built docopt
Installing collected packages: pymorphy2-dicts-ru, docopt, dawg-python, pymorphy2
Successfully installed dawg-python-0.7.2 docopt-0.6.2 pymorphy2-0.9.1 pymorphy2-dicts-ru-2.4.417127.4579844
```

```
import pandas as pd
import numpy as np
import gensim
import pymorphy2
from gensim import corpora
import nltk
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize
import re
```

<https://colab.research.google.com/drive/1HjEu9PnFkpvQfRfgPPzu5tjfudei1DUV?hl=ru&pli=1#scrollTo=VoAyEsiRL-yb&printMode=true>

3/19

14.03.2025, 08:25

Автоматический анализ сгенерированных диалогов средствами языка программирования Python.ipynb - Colab

```

# Загрузка необходимых ресурсов NLTK
nltk.download('punkt')
nltk.download('punkt_tab')
nltk.download('stopwords')

# Инициализация лемматизатора
morph = ru morphology2.MorphAnalyzer()

# Получение списка стоп-слов для русского языка
stop_words = set(stopwords.words('russian'))

# Добавление дополнительных стоп-слов
additional_stop_words = {
    'это', 'который', 'весь', 'наш', 'привет', 'собеседник',
    'это', 'этого', 'этому', 'этим', 'этом', 'свой', 'очень', 'какой', 'нужно', 'важно',
    'который', 'которая', 'которое', 'которые', 'которых', 'которому', 'которым', 'котором',
    'весь', 'вся', 'всё', 'всех', 'всему', 'всем', 'всей', 'всеми',
    'наш', 'наша', 'наше', 'наших', 'нашему', 'нашим', 'нашем', 'такой',
    'привет', 'привета', 'приветам', 'приветами', 'приветах',
    'собеседник', 'собеседника', 'собеседнику', 'собеседником', 'собеседниках',
    'и', 'а', 'но', 'или', 'да', 'что', 'чтобы', 'как', 'если', 'когда', 'хотя', 'потому что', 'либо', 'тем более',
    'в', 'на', 'с', 'по', 'к', 'за', 'от', 'до', 'перед', 'между', 'среди', 'о', 'про', 'ваш', 'мочь', 'хотеть', 'ты', 'ещё'
}
stop_words = stop_words.union(additional_stop_words)

# Функция для предобработки текста
def preprocess_text(text):
    # Удаление всех символов, кроме русских букв и пробелов
    text = re.sub(r'^а-яА-ЯЁё\s', '', text.lower())
    # Токенизация текста
    tokens = word_tokenize(text)
    # Фильтрация стоп-слов
    filtered_tokens = [word for word in tokens if word not in stop_words]
    # Лемматизация оставшихся слов
    return [morph.parse(word)[0].normal_form for word in filtered_tokens]

# Предобработка текста
df_sample['processed_text'] = df_sample['dialogue_text'].apply(preprocess_text)

# Создание словаря и корпуса
dictionary = corpora.Dictionary(df_sample['processed_text'])
corpus = [dictionary.doc2bow(text) for text in df_sample['processed_text']]

# Модель LDA
lda_model = gensim.models.LdaModel(corpus, num_topics=10, id2word=dictionary, passes=15)

# Вывод тем
for idx, topic in lda_model.print_topics(-1):
    print(f'Topic {idx}: {topic}')

```

 [nltk_data] Downloading package punkt to /root/nltk_data...
 [nltk_data] Unzipping tokenizers/punkt.zip.
 [nltk_data] Downloading package punkt_tab to /root/nltk_data...
 [nltk_data] Unzipping tokenizers/punkt_tab.zip.
 [nltk_data] Downloading package stopwords to /root/nltk_data...
 [nltk_data] Unzipping corpora/stopwords.zip.
 Topic 0: 0.012*"диалог" + 0.011*"продолжить" + 0.010*"место" + 0.009*"фильм" + 0.009*"любить" + 0.008*"девушка" + 0.008*"свой" + 0.017*"вопрос" + 0.018*"мочь" + 0.017*"консультант" + 0.017*"разный" + 0.017*"помогать" + 0.017*"умный" + 0.015*"парень" + 0.007*"помочь"
 Topic 1: 0.018*"вопрос" + 0.018*"мочь" + 0.017*"консультант" + 0.017*"разный" + 0.017*"помогать" + 0.017*"умный" + 0.015*"парень" + 0.007*"помочь"
 Topic 2: 0.032*"любить" + 0.017*"диалог" + 0.017*"продолжить" + 0.011*"увлечься" + 0.009*"свой" + 0.007*"заниматься" + 0.007*"раб
 Topic 3: 0.015*"диалог" + 0.013*"музыка" + 0.012*"какой" + 0.011*"любить" + 0.011*"продолжить" + 0.008*"нравиться" + 0.007*"свой" + 0.007*"язык"
 Topic 4: 0.014*"диалог" + 0.014*"продолжить" + 0.013*"любить" + 0.012*"увлечься" + 0.010*"время" + 0.010*"какой" + 0.010*"язык" + 0.010*"парень"
 Topic 5: 0.015*"мочь" + 0.015*"свой" + 0.012*"диалог" + 0.011*"продолжить" + 0.009*"новый" + 0.008*"хотеть" + 0.007*"помочь" + 0.007*"парень"
 Topic 6: 0.013*"мочь" + 0.012*"продолжить" + 0.012*"диалог" + 0.012*"вопрос" + 0.011*"человек" + 0.011*"любить" + 0.011*"парень" + 0.011*"парень"
 Topic 7: 0.027*"игра" + 0.013*"семья" + 0.011*"играть" + 0.010*"свой" + 0.009*"любить" + 0.008*"время" + 0.008*"друг" + 0.008*"прод
 Topic 8: 0.054*"любить" + 0.017*"работать" + 0.015*"диалог" + 0.015*"продолжить" + 0.011*"ребёнок" + 0.010*"заниматься" + 0.009*"хо
 Topic 9: 0.015*"мочь" + 0.013*"диалог" + 0.012*"продолжить" + 0.012*"вопрос" + 0.011*"разный" + 0.010*"парень" + 0.010*"помогать" +

Тематическое моделирование с использованием методов LDA и TF-IDF

```

from sklearn.feature_extraction.text import TfidfVectorizer

# Создание TF-IDF векторизатора
tfidf_vectorizer = TfidfVectorizer(tokenizer=lambda x: x.split(), lowercase=False)
tfidf_matrix = tfidf_vectorizer.fit_transform(df_sample['processed_text']).apply(lambda x: ' '.join(x))

# Применение LDA на TF-IDF матрице
from sklearn.decomposition import LatentDirichletAllocation

lda_model_tfidf = LatentDirichletAllocation(n_components=10, random_state=0)

```

<https://colab.research.google.com/drive/1HjEu9PnFkpvQfRfgPPz5tjfude1iDUV?hl=ru&pli=1#scrollTo=VoAyEsiRL-yb&printMode=true>

4/19

14.03.2025, 08:25

Автоматический анализ сгенерированных диалогов средствами языка программирования Python.ipynb - Colab

```
lda_model_tfidf.fit(tfidf_matrix)

# Вывод тем
for index, topic in enumerate(lda_model_tfidf.components_):
    print(f'Topic {index}:')
    print([tfidf_vectorizer.get_feature_names_out()[i] for i in topic.argsort()[-10:]])

/usr/local/lib/python3.10/dist-packages/sklearn/feature_extraction/text.py:517: UserWarning: The parameter 'token_pattern' will not
warnings.warn(
Topic 0:
['певец', 'канаеста', 'исполнитель', 'киноиндустрия', 'слава', 'поп', 'рок', 'цитата', 'сцена', 'инновация']
Topic 1:
['батарея', 'мигрень', 'собачка', 'букет', 'умереть', 'мораль', 'бальный', 'смешно', 'выносливость', 'пьяный']
Topic 2:
['помещение', 'вышивка', ' послушать', 'словно', 'бизнесвумен', 'налог', 'аккаунт', 'съёмка', 'экологический', 'минимум']
Topic 3:
['рис', 'обещание', 'президент', 'алкоголик', 'химия', 'волос', 'постановка', 'татуировщик', 'секс', 'ужасный']
Topic 4:
['готовить', 'продолжить', 'диалог', 'жить', 'звать', 'собака', 'заниматься', 'ребёнок', 'работать', 'любить']
Topic 5:
['системный', 'контент', 'секс', 'костюм', 'привлечь', 'вышивка', 'прыжок', 'парашот', 'оперный', 'маркетинг']
Topic 6:
['талантливый', 'медитировать', 'заявление', 'выступление', 'библиотекарь', 'кафе', 'парусный', 'выступить', 'дизайн', 'яхта']
Topic 7:
['шоколад', 'звезда', 'борода', 'инопланетянин', 'офицер', 'масло', 'остин', 'космос', 'астрономия', 'загадка']
Topic 8:
['новый', 'хотеть', 'помогать', 'вопрос', 'парень', 'продолжить', 'девушка', 'диалог', 'свой', 'мочь']
Topic 9:
['обама', 'придумать', 'жениться', 'программировать', 'сочи', 'фамилия', 'нейросетевой', 'куку', 'художница', 'автослесарь']
```

Тематическое моделирование с использованием преобученной модели BERT

```
from transformers import BertTokenizer, BertModel
import torch

# Загрузка модели и токенизатора BERT
tokenizer = BertTokenizer.from_pretrained('bert-base-multilingual-cased')
model = BertModel.from_pretrained('bert-base-multilingual-cased')

def get_embeddings(text):
    inputs = tokenizer(text, return_tensors='pt', truncation=True, padding=True)
    with torch.no_grad():
        outputs = model(**inputs)
    return outputs.last_hidden_state.mean(dim=1).squeeze().numpy()

# Применение к каждому диалогу
df_sample['embeddings'] = df_sample['dialogue_text'].apply(get_embeddings)

# Кластеризация
from sklearn.cluster import KMeans

# Применение KMeans к векторным представлениям
kmeans = KMeans(n_clusters=10, random_state=0)
df_sample['cluster'] = kmeans.fit_predict(list(df_sample['embeddings']))

# Анализ тематических кластеров
for cluster in range(10):
    print(f"\nКластер {cluster}:")
    print(df_sample[df_sample['cluster'] == cluster]['dialogue_text'].sample(3).tolist())
```

Уменьшение размерности эмбедингов для отображения текстов в двухмерном пространстве признаков

```
!pip install umap-learn
```

```
Collecting umap-learn
  Downloading umap_learn-0.5.7-py3-none-any.whl.metadata (21 kB)
Requirement already satisfied: numpy>=1.17 in /usr/local/lib/python3.10/dist-packages (from umap-learn) (1.26.4)
Requirement already satisfied: scipy>=1.3.1 in /usr/local/lib/python3.10/dist-packages (from umap-learn) (1.13.1)
Requirement already satisfied: scikit-learn>=0.22 in /usr/local/lib/python3.10/dist-packages (from umap-learn) (1.6.0)
Requirement already satisfied: numba>=0.51.2 in /usr/local/lib/python3.10/dist-packages (from umap-learn) (0.60.0)
Collecting pynndescent>=0.5 (from umap-learn)
  Downloading pynndescent-0.5.13-py3-none-any.whl.metadata (6.8 kB)
Requirement already satisfied: tqdm in /usr/local/lib/python3.10/dist-packages (from umap-learn) (4.67.1)
Requirement already satisfied: llvmlite<0.44,>=0.43.0dev0 in /usr/local/lib/python3.10/dist-packages (from numba>=0.51.2->umap-learn) (0.43.0)
Requirement already satisfied: joblib>=0.11 in /usr/local/lib/python3.10/dist-packages (from pynndescent>=0.5->umap-learn) (1.4.2)
Requirement already satisfied: threadpoolctl>=3.1.0 in /usr/local/lib/python3.10/dist-packages (from scikit-learn>=0.22->umap-learn) (3.3.0)
Downloading umap_learn-0.5.7-py3-none-any.whl (88 kB)
----- 88.8/88.8 kB 6.6 MB/s eta 0:00:00
```

14.03.2025, 08:25

Автоматический анализ сгенерированных диалогов средствами языка программирования Python.ipynb - Colab

```

Downloading pynndescent-0.5.13-py3-none-any.whl (56 kB)
56.9/56.9 kB 5.5 MB/s eta 0:00:00
Installing collected packages: pynndescent, umap-learn
Successfully installed pynndescent-0.5.13 umap-learn-0.5.7

```

```

import umap.umap_ as umap
import matplotlib.pyplot as plt

# Снижение размерности с помощью UMAP
umap_model = umap.UMAP(n_neighbors=5, n_components=2, metric='cosine')
umap_embeddings = umap_model.fit_transform(list(df_sample['embeddings']))

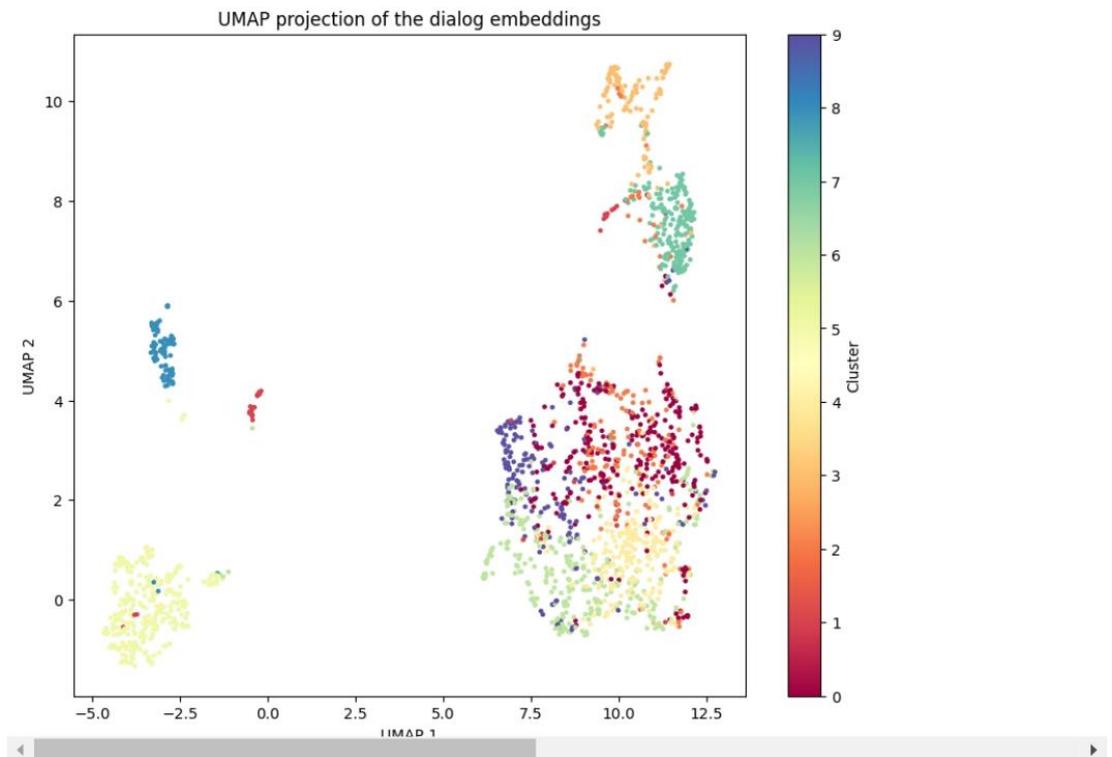
# Визуализация
plt.figure(figsize=(10, 8))
scatter = plt.scatter(umap_embeddings[:, 0], umap_embeddings[:, 1], c=df_sample['cluster'], cmap='Spectral', s=5)
plt.title('UMAP projection of the dialog embeddings')
plt.colorbar(scatter, label='Cluster')
plt.xlabel('UMAP 1')
plt.ylabel('UMAP 2')
plt.show()

```

```

/usr/local/lib/python3.10/dist-packages/sklearn/utils/deprecation.py:151: FutureWarning: 'force_all_finite' was renamed to 'ensure_
warnings.warn(

```



Тематическое моделирование с обучением на предварительно размеченных текстах

```

df_sample.to_csv('df_sample.txt', sep='\t', index=False)

import zipfile
import xml.etree.ElementTree as ET
import os

zip_file_path = '/content/ДИАЛОГИ.zip'

os.makedirs('/content/dialogs', exist_ok=True)

with zipfile.ZipFile(zip_file_path, 'r') as zip_ref:
    zip_ref.extractall('/content/dialogs')

```

<https://colab.research.google.com/drive/1HJEu9PnFkpvQfRfgPPzu5tjfade1iDUV?hl=ru&pli=1#scrollTo=VoAyEsiRL-yb&printMode=true>

6/19

14.03.2025, 08:25

Автоматический анализ сгенерированных диалогов средствами языка программирования Python.ipynb - Colab

```

data = []

# Сортировка списка файлов
file_list = sorted(os.listdir('/content/dialogs'), key=lambda x: int(x.split('.')[0]))

# Проходим по всем файлам в папке
for filename in file_list:
    if filename.endswith('.txt'):
        # Чтение файла
        file_path = os.path.join('/content/dialogs', filename)
        print(f"Чтение файла: {file_path}")
        try:
            tree = ET.parse(file_path)
            root = tree.getroot()

            # Извлечение темы диалога
            topic = root.find('topic').text if root.find('topic') is not None else None

            # Извлечение высказываний
            for turn in root.findall('turn'):
                speaker = turn.find('speaker').text if turn.find('speaker') is not None else None
                utterance = turn.find('utterance').text if turn.find('utterance') is not None else ""
                data.append({'topic': topic, 'speaker': speaker, 'utterance': utterance})

        except Exception as e:
            print(f"Ошибка при обработке файла {filename}: {e}")

# Создание DataFrame из списка данных
df = pd.DataFrame(data)

# Вывод DataFrame
df

```

```

Чтение файла: /content/dialogs/1.txt
Чтение файла: /content/dialogs/2.txt
Чтение файла: /content/dialogs/3.txt
Чтение файла: /content/dialogs/4.txt
Чтение файла: /content/dialogs/5.txt
Чтение файла: /content/dialogs/6.txt
Чтение файла: /content/dialogs/7.txt
Чтение файла: /content/dialogs/8.txt
Чтение файла: /content/dialogs/9.txt
Чтение файла: /content/dialogs/10.txt
Чтение файла: /content/dialogs/11.txt
Чтение файла: /content/dialogs/12.txt
Чтение файла: /content/dialogs/13.txt
Чтение файла: /content/dialogs/14.txt
Чтение файла: /content/dialogs/15.txt
Чтение файла: /content/dialogs/16.txt
Чтение файла: /content/dialogs/17.txt
Чтение файла: /content/dialogs/18.txt
Чтение файла: /content/dialogs/19.txt
Чтение файла: /content/dialogs/20.txt
Ошибка при обработке файла 20.txt: syntax error: line 1, column 0

```

	topic	speaker	utterance
0	Запрос информации	Собеседник	Я загрузил архив на Яндекс Диск, но он закачал...
1	Запрос информации	Ты	Конечно, есть несколько способов восстановлени...
2	Запрос информации	Собеседник	Как я могу проверить, был ли файл загружен пол...
3	Запрос информации	Ты	Вы можете проверить размер файла на вашем комп...
4	Запрос информации	Собеседник	А как я могу воспользоваться функцией Версии н...
...
139	Запрос информации	Ты	Да, суд может взыскать расходы, связанные с пр...
140	Запрос информации	Собеседник	Что такое интернет-пиратство и чем оно отличае...
141	Запрос информации	Ты	Есть интересная параллель с тем, о чем вы здес...
142	Запрос информации	Собеседник	Как это сходит с рук интернет-пиратам?
143	Запрос информации	Ты	К сожалению, у людей есть много способов делат...

```
df['topic'].value_counts()
```

14.03.2025, 08:25

Автоматический анализ сгенерированных диалогов средствами языка программирования Python.ipynb - Colab



topic	count
Запрос информации	38
Знакомство	36
Домашние животные	28
Спорт и активность	12
Домашние дела	12
Хобби	8
Поиск работы	6
Саморазвитие и личностный рост	4

Извлечение реплик каждого спикера из диалогов

```
data = []

input_file = '/content/df_sample.txt'

# Регулярное выражение для поиска спикеров и их реплик
dialog_pattern = re.compile(r'(?P<speaker>Собеседник|Ты): (?P<utterance>.+?)(?=\n\S+|$)', re.DOTALL)
print(f"Чтение файла: {input_file}")
try:
    with open(input_file, 'r', encoding='utf-8') as f:
        content = f.read()
        matches = dialog_pattern.finditer(content)
        for match in matches:
            speaker = match.group('speaker')
            utterance = match.group('utterance').strip() # Удаляем лишние пробелы
            data.append({'speaker': speaker, 'utterance': utterance})
except Exception as e:
    print(f"Ошибка при обработке файла: {e}")
df1 = pd.DataFrame(data)
df1
```



Чтение файла: /content/df_sample.txt

	speaker	utterance
0	Собеседник	Я задумался о Джордже Вашингтоне и его действи...
1	Ты	Действительно, Джордж Вашингтон нарушил присяг...
2	Собеседник	Однако, когда корона начала ущемлять права аме...
3	Ты	Кроме того, важно учитывать, что американская ...
4	Собеседник	Но почему же Вашингтон не получил каких-то юри...
...
16764	Ты	Привет! Меня Алевтина! Как твои дела? Я очень ...
16765	Собеседник	А я не люблю читать, я люблю рисовать. Посовет...
16766	Ты	Ну ничего, ты молодец, мои внуки тоже рисуют, ...
16767	Собеседник	Я люблю кушать суши.
16768	Ты	Супер, а мы любим пироги есть которые я готовл...

Обучение тематическому моделированию на размеченных данных с помощью простой логистической регрессии

```
!pip install pymorphy2

Requirement already satisfied: pymorphy2 in /usr/local/lib/python3.10/dist-packages (0.9.1)
Requirement already satisfied: dawg-python>=0.7.1 in /usr/local/lib/python3.10/dist-packages (from pymorphy2) (0.7.2)
Requirement already satisfied: pymorphy2-dicts-ru<3.0,>=2.4 in /usr/local/lib/python3.10/dist-packages (from pymorphy2) (2.4.417127)
Requirement already satisfied: docopt>=0.6 in /usr/local/lib/python3.10/dist-packages (from pymorphy2) (0.6.2)

!pip install nltk
```

14.03.2025, 08:25

Автоматический анализ сгенерированных диалогов средствами языка программирования Python.ipynb - Colab

```

Requirement already satisfied: nltk in /usr/local/lib/python3.10/dist-packages (3.9.1)
Requirement already satisfied: click in /usr/local/lib/python3.10/dist-packages (from nltk) (8.1.8)
Requirement already satisfied: joblib in /usr/local/lib/python3.10/dist-packages (from nltk) (1.4.2)
Requirement already satisfied: regex>=2021.8.3 in /usr/local/lib/python3.10/dist-packages (from nltk) (2024.11.6)
Requirement already satisfied: tqdm in /usr/local/lib/python3.10/dist-packages (from nltk) (4.67.1)

```

```

import pandas as pd
import re
from sklearn.model_selection import train_test_split
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import classification_report

# Пример обучающего набора данных
train_df = df

# Пример нового набора данных
new_df = df1

# Предобработка текста
def preprocess_text(text):
    text = re.sub(r'^а-яА-ЯЁё\s', '', text)
    return text.lower()

train_df['processed_text'] = train_df['utterance'].apply(preprocess_text)
new_df['processed_text'] = new_df['utterance'].apply(preprocess_text)

# Векторизация текста
vectorizer = TfidfVectorizer()
X_train = vectorizer.fit_transform(train_df['processed_text'])
y_train = train_df['topic']

# Обучение модели
model = LogisticRegression()
model.fit(X_train, y_train)

# Векторизация нового набора данных
X_new = vectorizer.transform(new_df['processed_text'])

# Предсказание тем
new_df['predicted_topic'] = model.predict(X_new)

# Вывод результатов
new_df[['speaker', 'utterance', 'predicted_topic']]

```

	speaker	utterance	predicted_topic
0	Собеседник	Я задумался о Джордже Вашингтоне и его действи...	Знакомство
1	Ты	Действительно, Джордж Вашингтон нарушил присяг...	Запрос информации
2	Собеседник	Однако, когда корона начала ущемлять права аме...	Запрос информации
3	Ты	Кроме того, важно учитывать, что американская ...	Запрос информации
4	Собеседник	Но почему же Вашингтон не получил каких-то юри...	Знакомство
...
16764	Ты	Привет! Меня Алевтина! Как твои дела? Я очень ...	Знакомство
16765	Собеседник	А я не люблю читать, я люблю рисовать. Посовет...	Знакомство
16766	Ты	Ну ничего, ты молодец, мои внуки тоже рисуют, ...	Знакомство
16767	Собеседник	Я люблю кушать суши.	Знакомство
16768	Ты	Супер, а мы любим пироги есть которые я готовл...	Знакомство
16769	Ты

Оценка результатов тематического моделирования с помощью простой логистической регрессии (распределение диалогов по темам)

```
new_df['predicted_topic'].value_counts()
```

14.03.2025, 08:25

Автоматический анализ сгенерированных диалогов средствами языка программирования Python.ipynb - Colab

	count
predicted_topic	
Запрос информации	9512
Знакомство	6168
Домашние животные	1088
Спорт и активность	1

Обучение тематической разметке с использованием метода "случайный лес" и TfidfVectorizer (преобразование текстовых данных в векторные представления с использованием метода TF-IDF)

```
import pandas as pd
import re
from sklearn.model_selection import train_test_split, cross_val_score
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import classification_report
from sklearn.pipeline import Pipeline
from nltk.corpus import stopwords
from nltk.stem import WordNetLemmatizer
import nltk

# Загрузка необходимых ресурсов для NLTK
nltk.download('stopwords')
nltk.download('wordnet')

# Предобработка текста
def preprocess_text(text):
    lemmatizer = WordNetLemmatizer()
    stop_words = set(stopwords.words('russian'))
    text = re.sub(r'^а-яА-ЯЁЁ\s', '', text)
    text = text.lower()
    text = ' '.join([lemmatizer.lemmatize(word) for word in text.split() if word not in stop_words])
    return text

train_df['processed_text'] = train_df['utterance'].apply(preprocess_text)
new_df['processed_text'] = new_df['utterance'].apply(preprocess_text)

# Создание пайплайна с векторизацией и моделью
pipeline = Pipeline([
    ('tfidf', TfidfVectorizer()),
    ('classifier', RandomForestClassifier(n_estimators=100, random_state=42))
])

# Обучение модели с кросс-валидацией
X_train = train_df['processed_text']
y_train = train_df['topic']
pipeline.fit(X_train, y_train)

# Оценка модели с кросс-валидацией
cv_scores = cross_val_score(pipeline, X_train, y_train, cv=5)
print(f'Кросс-валидация: Средняя точность = {cv_scores.mean():.4f}')

# Предсказание тем для нового набора данных
new_df['predicted_topic'] = pipeline.predict(new_df['processed_text'])

# Вывод результатов
new_df[['speaker', 'utterance', 'predicted_topic']]
```

```
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data] Package stopwords is already up-to-date!
[nltk_data] Downloading package wordnet to /root/nltk_data...
/usr/local/lib/python3.10/dist-packages/sklearn/model_selection/_split.py:805: UserWarning: The least populated class in y has only
warnings.warn(
Кросс-валидация: Средняя точность = 0.3057
```

	speaker	utterance	predicted_topic
0	Собеседник	Я задумался о Джордже Вашингтоне и его действи...	Знакомство
1	Ты	Действительно, Джордж Вашингтон нарушил присяг...	Знакомство
2	Собеседник	Однако, когда корона начала ущемлять права аме...	Знакомство
3	Ты	Кроме того, важно учитывать, что американская ...	Знакомство
4	Собеседник	Но почему же Вашингтон не получил каких-то юри...	Знакомство
...
16764	Ты	Привет! Меня Алевтина! Как твои дела? Я очень ...	Домашние дела
16765	Собеседник	А я не люблю читать, я люблю рисовать. Посовет...	Знакомство
16766	Ты	Ну ничего, ты молодец, мои внуки тоже рисуют, ...	Домашние дела
16767	Собеседник	Я люблю кушать суши.	Знакомство
16768	Ты	Супер, а мы любим пироги есть которые я готовл...	Знакомство

Оценка результатов обучения с использованием метода "случайный лес"

```
new_df['predicted_topic'].value_counts()
```

predicted_topic	count
Знакомство	14312
Домашние животные	1317
Запрос информации	514
Домашние дела	346
Спорт и активность	130
Саморазвитие и личностный рост	88
Хобби	48
Поиск работы	14

```
new_df
```

	speaker	utterance	processed_text	predicted_topic
0	Собеседник	Я задумался о Джордже Вашингтоне и его действи...	задумался джордже вашингтоне действиях время р...	Знакомство
1	Ты	Действительно, Джордж Вашингтон нарушил присяг...	действительно джордж вашингтон нарушил присягу...	Знакомство
2	Собеседник	Однако, когда корона начала ущемлять права аме...	однако корона начала ущемлять права американце...	Знакомство
3	Ты	Кроме того, важно учитывать, что американская ...	кроме важно учитывать американская революция Д...	Знакомство
4	Собеседник	Но почему же Вашингтон не получил каких-то юри...	почему вашингтон получил какихто юридических п...	Знакомство
...
16764	Ты	Привет! Меня Алевтина! Как твои дела? Я очень ...	привет алевтина твои дела очень люблю	Домашние дела
16765	Собеседник	А я не люблю читать, я люблю рисовать. Посовет...	люблю читать люблю рисовать посоветовать выбор...	Знакомство

14.03.2025, 08:25

Автоматический анализ сгенерированных диалогов средствами языка программирования Python.ipynb - Colab

!python -m spacy download ru_core_news_sm

 Показать скрытые выходные данные

Вычисление текстовых метрик (в разрезе тем для каждого спикера)

```

import numpy as np
import pandas as pd
from collections import Counter

def calculate_metrics(group):
    metrics = {}

    # Средняя длина реплик
    metrics['average_utterance_length'] = group['processed_text'].apply(lambda x: len(x.split())).mean()

    # Средняя длина предложений
    metrics['average_sentence_length'] = group['utterance'].apply(lambda x: len(x.split('.')).mean())

    # Индекс лексического разнообразия
    total_words = group['processed_text'].str.split().apply(len).sum()
    unique_words = len(set(' '.join(group['processed_text']).split()))
    metrics['lexical_diversity_index'] = unique_words / total_words if total_words > 0 else 0

    # Индекс Флеша (упрощенный)
    metrics['flesch_index'] = 206.835 - 1.015 * (total_words / group['processed_text'].count()) - 84.6 * (len(group['processed_text'].split()))

    # Доля вопросительных предложений
    metrics['question_percentage'] = (group['utterance'].str.count(r'\?').sum() / group.shape[0]) * 100

    # Доля восклицательных предложений
    metrics['exclamation_percentage'] = (group['utterance'].str.count(r'!').sum() / group.shape[0]) * 100

    # Доля повествовательных предложений
    metrics['narrative_percentage'] = 100 - (metrics['question_percentage'] + metrics['exclamation_percentage'])

    return pd.Series(metrics)

# Группировка данных по спикерам и темам
grouped = new_df.groupby(['speaker', 'predicted_topic'])

# Расчет метрик
metrics_df = grouped.apply(calculate_metrics).reset_index()

```

 <ipython-input-43-7b39f63865a3>:37: DeprecationWarning: DataFrameGroupBy.apply operated on the grouping columns. This behavior is deprecated. Use metrics_df = grouped.apply(calculate_metrics).reset_index()

metrics_df

14.03.2025, 08:25

Автоматический анализ сгенерированных диалогов средствами языка программирования Python.ipynb - Colab

	speaker	predicted_topic	average_utterance_length	average_sentence_length	lexical_diversity_index	flesch_index	question
0	Собеседник	Домашние дела	4.642857	1.904762	0.607692	117.522500	
1	Собеседник	Домашние животные	8.712891	2.304688	0.530150	113.391416	
2	Собеседник	Запрос информации	13.209302	2.209302	0.663732	108.827558	
3	Собеседник	Знакомство	5.495998	1.840315	0.303704	116.656562	
4	Собеседник	Поиск работы	9.000000	2.000000	1.000000	113.100000	
5	Собеседник	Саморазвитие и личностный рост	10.935484	2.387097	0.719764	111.135484	
6	Собеседник	Спорт и активность	7.600000	2.054545	0.717703	114.521000	
7	Собеседник	Хобби	1.565217	1.869565	0.388889	120.646304	
8	Ты	Домашние дела	12.612360	2.140449	0.502004	109.433455	
9	Ты	Домашние животные	45.336646	2.680745	0.243972	76.218304	
10	Ты	Запрос информации	79.577103	3.474299	0.254940	41.464241	
11	Ты	Знакомство	22.119865	2.388351	0.159754	99.783337	

Изучение структуры диалогов путем анализа перехода между репликами

```
# Создание столбца с предыдущей репликой
new_df['previous_utterance'] = new_df['utterance'].shift(1)
# Создание столбца с переходами
new_df['transition'] = new_df['previous_utterance'].astype(str) + ' -> ' + new_df['utterance'].astype(str)
# Подсчет переходов
transition_counts = new_df['transition'].value_counts()

# Выбор топ-N переходов для визуализации
top_n = transition_counts
# Вывод таблицы
table = pd.DataFrame(top_n.reset_index())
table.columns = ['Переход', 'Частота']
```

table

	Переход	Частота
0	Привет. -> Привет.	49
1	привет. -> Привет.	10
2	Привет. -> Как дела?	9
3	Привет! -> Привет.	9
4	привет -> здравствуй	5
...
12911	Недавно мы устроили караоке-баттл и ни один го...	1
12912	Утром всегда все идет не по плану, постоянная ...	1
12913	О, не переживай, утро - это только начало дня!...	1
12914	Я просто не успеваю сделать все, что запланиро...	1
12915	Каковы распространенные причины сухости во рту...	1

12916 rows x 2 columns

Автоматическая классификация реплик по типам переходов

```
def classify_transition(utterance):
    # Ключевые слова для каждой категории
    greetings = ['привет', 'приветик', 'здравствуй', 'доброе утро', 'добрый вечер', 'приветствую']
    info_requests = ['как', 'что', 'где', 'когда', 'почему', 'зачем', 'интересно', 'сколько', 'ты знаешь', 'ты не знаешь', 'могу', 'расскажи']
    comments = ['думаю', 'считаю', 'мне кажется', 'по моему мнению', 'говоря о', 'знаешь', 'хорошо', 'послушай', 'слушай']
    answers = ['да', 'действительно', 'нет', 'можете', 'попробуй', 'попробуйте', 'проверьте', 'проверь', 'посмотри', 'правильно', 'неправильно']
    emotional_evaluation = ['вау', 'вауу', 'ваууу', 'вауууу', 'ваууууу', 'интересно', 'интересно!', 'интересно!!', 'интересно!!!', 'интересно!!!!', 'интересно!!!!!']

https://colab.research.google.com/drive/1HjEu9PnFkpvQfRfgPPZu5tjfude1iDUV?hl=ru&pli=1#scrollTo=VoAyEsiRL-yb&printMode=true 13/19
```

14.03.2025, 08:25

Автоматический анализ сгенерированных диалогов средствами языка программирования Python.ipynb - Colab

```

emotional_evaluations = ['рад', 'класс', 'классно', 'удивлен', 'хорошо', 'интересно', 'к сожалению', 'здорово', 'удивительно', 'оо', 'с
info_messages = ['у меня', 'для меня', 'работаю', 'женат', 'жена', 'муж', 'дочь', 'сын', 'замужем', 'люблю', 'живу', 'рисую', 'читаю', 'этикетные реплики']
etiquette_messages = ['спасибо', 'благодарю', 'просьба', 'взаимно', 'пожалуйста', 'рад помочь', 'с удовольствием', 'не за что', 'всег

utterance_lower = utterance.lower()

if any(keyword in utterance_lower for keyword in greetings):
    return 'приветствия'
elif any(keyword in utterance_lower for keyword in info_requests):
    return 'запросы информации'
elif any(keyword in utterance_lower for keyword in comments):
    return 'комментарии'
elif any(keyword in utterance_lower for keyword in answers):
    return 'ответ на вопрос'
elif any(keyword in utterance_lower for keyword in emotional_evaluations):
    return 'эмоционально-оценочная реплика'
elif any(keyword in utterance_lower for keyword in info_messages):
    return 'сообщение информации'
elif any(keyword in utterance_lower for keyword in etiquette_messages):
    return 'этикетные реплики'
else:
    return 'сообщение информации'

# Применение классификации к репликам
new_df['transition_category'] = new_df['utterance'].apply(classify_transition)

# Подсчет переходов по категориям
new_df['next_speaker'] = new_df['speaker'].shift(-1)
new_df['next_utterance'] = new_df['utterance'].shift(-1)

# Фильтрация переходов
transitions = new_df[['speaker', 'utterance', 'next_speaker', 'next_utterance', 'transition_category']].dropna()

# Подсчет частоты переходов по категориям
transition_counts = transitions.groupby(['transition_category']).size().reset_index(name='count')

# Сортировка по частоте
transition_counts = transition_counts.sort_values(by='count', ascending=False)

print(transition_counts)

```

	transition_category	count
0	запросы информации	8553
4	сообщение информации	3659
2	ответ на вопрос	2348
3	приветствия	1200
5	эмоционально-оценочная реплика	661
1	комментарии	274
6	этикетные реплики	73

Результаты анализа структуры диалогов через исследование речевых паттернов смены спикеров (переходов между репликами)

```
new_df
```

14.03.2025, 08:25

Автоматический анализ сгенерированных диалогов средствами языка программирования Python.ipynb - Colab

	speaker	utterance	processed_text	predicted_topic	previous_utterance	transition	transition_category	next_s
0	Собеседник	Я задумался о Джордже Вашингтоне и его действи...	задумался джордже вашингтоне действиях время р...	Знакомство	None	None -> Я задумался о Джордже Вашингтоне и его...	запросы информации	
1	Ты	Действительно, Джордж Вашингтон нарушил присяг...	действительно джордж вашингтон нарушил присягу...	Знакомство	Я задумался о Джордже Вашингтоне и его действи...	Я задумался о Джордже Вашингтоне и его действи...	запросы информации	Собес
2	Собеседник	Однако, когда корона начала ущемлять права аме...	однако корона начала ущемлять права американце...	Знакомство	Действительно, Джордж Вашингтон нарушил присяг...	Действительно, Джордж Вашингтон нарушил присяг...	запросы информации	
3	Ты	Кроме того, важно учитывать, что американская ...	кроме важно учитывать американская революция д...	Знакомство	Однако, когда корона начала ущемлять права аме...	Однако, когда корона начала ущемлять права аме...	запросы информации	Собес
4	Собеседник	Но почему же Вашингтон не получил каких-то юри...	почему вашингтон получил какихто юридических п...	Знакомство	Кроме того, важно учитывать, что американская ...	Кроме того, важно учитывать, что американская ...	запросы информации	
...

Визуализация полученных структурных элементов (категорий)

```

import matplotlib.pyplot as plt
import seaborn as sns

# Загрузка данных
data = new_df

# Создание колонки для следующего говорящего и следующей реплики
data['next_speaker'] = data['speaker'].shift(-1)
data['next_utterance'] = data['utterance'].shift(-1)

# Фильтрация переходов
transitions = data[['predicted_topic', 'speaker', 'utterance', 'next_speaker', 'next_utterance', 'transition_category']].dropna()

# Подсчет переходов по темам
transition_counts = transitions.groupby(['predicted_topic', 'speaker', 'next_speaker']).size().reset_index(name='count')

# Визуализация переходов
plt.figure(figsize=(12, 6))
sns.barplot(data=transition_counts, x='predicted_topic', y='count', hue='speaker')
plt.title('Переходы между репликами по темам')
plt.xlabel('Темы')
plt.ylabel('Количество переходов')
plt.legend(title='Говорящий')
plt.xticks(rotation=45)
plt.tight_layout()
plt.show()

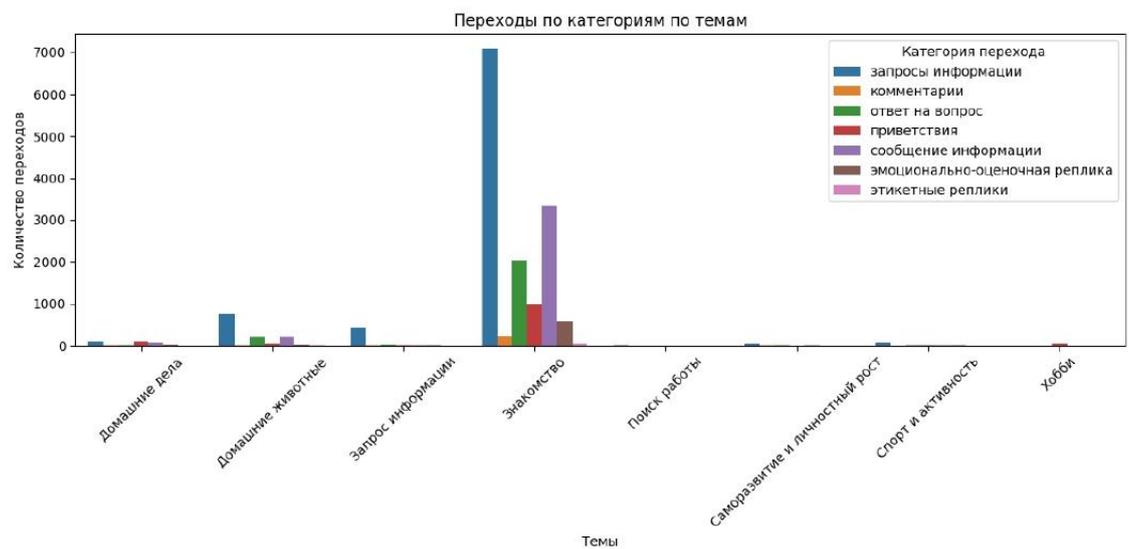
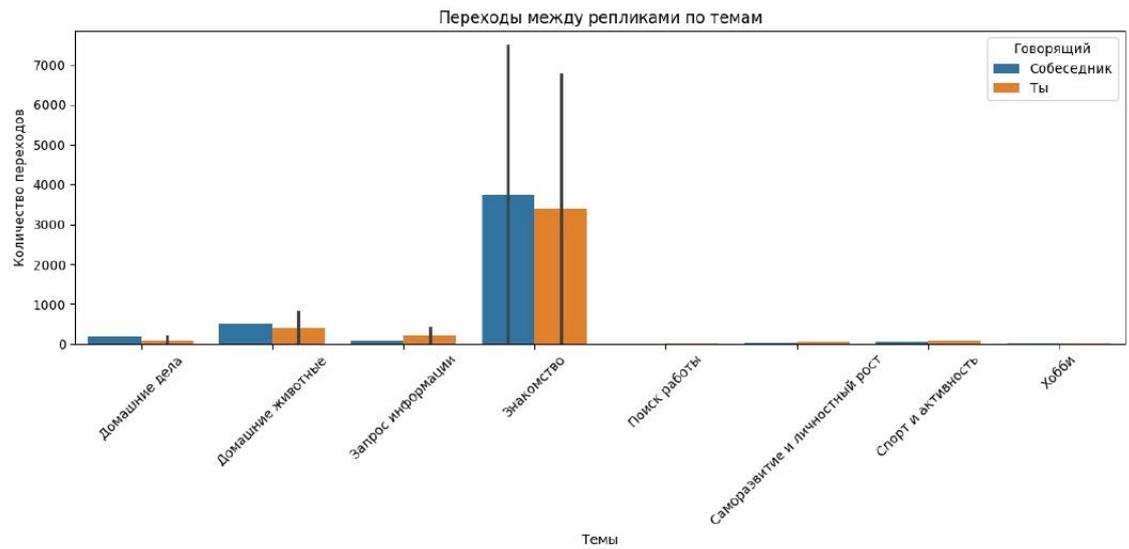
# Подсчет переходов по категориям
transition_category_counts = transitions.groupby(['predicted_topic', 'transition_category']).size().reset_index(name='count')

# Визуализация переходов по категориям
plt.figure(figsize=(12, 6))
sns.barplot(data=transition_category_counts, x='predicted_topic', y='count', hue='transition_category')
plt.title('Переходы по категориям по темам')
plt.xlabel('Темы')
plt.ylabel('Количество переходов')
plt.legend(title='Категория перехода')
plt.xticks(rotation=45)
plt.tight_layout()
plt.show()

```

14.03.2025, 08:25

Автоматический анализ сгенерированных диалогов средствами языка программирования Python.ipynb - Colab



Морфолого-синтаксический анализ текстов диалогов

```
#import pandas as pd
#import pymorphy2
import spacy

# Загружаем модель для русского языка
nlp = spacy.load('ru_core_news_sm')

# Инициализация морфологического анализатора
morph = pymorphy2.MorphAnalyzer()

# Функция для разметки текста
def annotate_text(text):
    words = text.split()
    annotated_words = []
```

<https://colab.research.google.com/drive/1HjEu9PnFkpVQfRfgPPZu5tjfude1iDUV?hl=ru&pli=1#scrollTo=VoAyEsiRL-yb&printMode=true>

16/19

14.03.2025, 08:25 Автоматический анализ сгенерированных диалогов средствами языка программирования Python.ipynb - Colab

```
# Выполнение синтаксического анализа с помощью spaCy
doc = nlp(text)

for token in doc:
    word_clean = token.text
    normal_form = morph.parse(word_clean)[0].normal_form # Нормальная форма
    pos = token.pos_ # Часть речи из spaCy
    tag = token.tag_ # Синтаксическая информация из spaCy

    annotated_words.append((word_clean, normal_form, pos, tag))

return annotated_words

# Разметка текста в столбце
new_df['annotated'] = new_df['utterance'].apply(annotate_text)

new_df['annotated']
```

	annotated
0	[(Я, я, PRON, PRON), (задумался, задуматься, V...
1	[(Действительно, действительно, ADV, ADV), (...
2	[(Однако, однако, ADV, ADV), (,, PUNCT, PUN...
3	[(Кроме, кроме, ADP, ADP), (того, тот, PRON, P...
4	[(Но, но, CCONJ, CCONJ), (почему, почему, ADV...
...	...
16764	[(Привет, привет, NOUN, NOUN), (,, PUNCT, P...
16765	[(А, а, CCONJ, CCONJ), (я, я, PRON, PRON), (не...
16766	[(Ну, ну, PART, PART), (ничего, ничего, PRON, ...
16767	[(Я, я, PRON, PRON), (люблю, любить, VERB, VER...
16768	[(Супер, супер, NOUN, NOUN), (,, PUNCT, PUN...

16769 rows x 1 columns

```
# Подсчет частей речи
pos_counts = {}

for annotations in new_df['annotated']:
    for _, _, pos, _ in annotations:
        if pos in pos_counts:
            pos_counts[pos] += 1
        else:
            pos_counts[pos] = 1

print("Подсчет частей речи:")
print(pos_counts)
```

```
Подсчет частей речи:
{'PRON': 28002, 'VERB': 79724, 'ADP': 20217, 'PROPN': 6697, 'CCONJ': 13119, 'DET': 14458, 'NOUN': 115012, 'PUNCT': 496639, 'ADV': 2...
```

```
# Функция для извлечения синтаксических зависимостей
def extract_syntax_dependencies(text):
    doc = nlp(text)
    dependencies = []

    for token in doc:
        dependencies.append({
            'word': token.text,
            'lemma': token.lemma_,
            'pos': token.pos_,
            'dep': token.dep_,
            'head': token.head.text,
            'head_pos': token.head.pos_,
            'subtree': [child.text for child in token.children] # Дочерние элементы
        })

    return dependencies

new_df['syntax_dependencies'] = new_df['utterance'].apply(extract_syntax_dependencies)
```

14.03.2025, 08:25

Автоматический анализ сгенерированных диалогов средствами языка программирования Python.ipynb - Colab

```
# Выводим результаты
# for index, row in new_df.iterrows():
#     print(f"Текст: {row['utterance']}")
#     for dep in row['syntax_dependencies']:
#         print(f" Слово: {dep['word']}, Лемма: {dep['lemma']}, POS: {dep['pos']}, "
#               f"Зависимость: {dep['dep']}, Глава: {dep['head']}, Глава POS: {dep['head_pos']}, "
#               f"Поддереву: {dep['subtree']}")
#     print("\n")
```

```
⚠ /usr/local/lib/python3.10/dist-packages/spacy/util.py:1740: UserWarning: [W111] Jupyter notebook detected: if using `prefer_gpu()` ,
warnings.warn(Warnings.W111)
```

```
# Подсчет частоты частей речи
pos_counts = new_df['syntax_dependencies'].apply(lambda deps: [dep['pos'] for dep in deps]).explode().value_counts()

# Подсчет частоты зависимостей
dep_counts = new_df['syntax_dependencies'].apply(lambda deps: [dep['dep'] for dep in deps]).explode().value_counts()

# Создание DataFrame для визуализации
pos_summary = pos_counts.reset_index()
pos_summary.columns = ['Часть речи', 'Количество']

dep_summary = dep_counts.reset_index()
dep_summary.columns = ['Зависимость', 'Количество']

print("Частота частей речи:")
print(pos_summary)

print("\nЧастота зависимостей:")
print(dep_summary)
```

```
↔ Частота частей речи:
```

Часть речи	Количество	
0	PUNCT	496639
1	NOUN	115012
2	VERB	79724
3	ADJ	44213
4	PRON	28002
5	ADV	25388
6	ADP	20217
7	DET	14458
8	CCONJ	13119
9	PART	8703
10	PROPN	6697
11	SCONJ	5964
12	NUM	2882
13	AUX	2545
14	SPACE	2001
15	X	1896
16	INTJ	734
17	SYM	4

```
Частота зависимостей:
```

Зависимость	Количество	
0	punct	496731
1	conj	132271
2	ROOT	30794
3	nsubj	26478
4	case	20676
5	advmod	20341
6	parataxis	18066
7	amod	17170
8	obl	14961
9	obj	14568
10	cc	13102
11	det	12543
12	nmod	12291
13	xcomp	7948
14	mark	5112
15	csubj	3315
16	iobj	3095
17	appos	2585
18	advcl	2161
19	ccomp	2159
20	dep	2001
21	fixed	1561
22	cop	1449
23	acl:relcl	1432
24	acl	1069
25	nummod:gov	830
26	aux	781
27	nsubj:pass	734
28	nummod	588

Синтетические диалоги для атрибуции признаков генерации

Диалог 1: Гарри Поттер

Аня: Слушай, Гарри – он вообще нормальный тип? Иногда такой тормоз, меня бесит.

Дима: Ну, он же сирота, жизнь не сахар. Представь себя на его месте!

Аня: Да понимаю я, просто иногда хочется его встряхнуть! Ну почему он такой наивный?

Дима: Он доверяет людям. Это круто.

Аня: Ну доверял бы хоть тем, кому надо! А то вечно влипает.

Дима: Зато спасает мир! Герой.

Диалог 2: Гермиона Грейнджер

Лена: Гермиона – заучка, но она классная. Без неё Гарри бы давно загнулся.

Сергей: Ага, мозг команды. И храбрая к тому же.

Лена: Да, и знаешь, она же борется за справедливость. Эльфы там, всё такое.

Сергей: Это да, она прямо активистка.

Лена: И ведь она такая же, как мы.

Сергей: Ага.

Диалог 3: Рон Уизли

Игорь: Рон – прикольный. Всегда голодный и боится пауков.

Ольга: Зато верный друг. И смешной.

Игорь: Да, его шутки иногда просто огонь.

Ольга: И он растет, он меняется.

Диалог 4: Невилл Лонгботтом

Катя: Невилл – сначала такой неуклюжий, а потом как расцвёл!

Павел: Да, герой! В конце вообще круто сражался.

Катя: Меня он больше всего зацепил.

Диалог 5: Луна Лавгуд

Соня: Луна – такая странная...

Максим: Ну и что? Она зато искренняя и не боится быть собой.

Соня: Да, она как будто из другого мира.

Максим: Наверное поэтому она такая классная.

Диалог 6: Драко Малфой

Вика: Драко – он вообще кто? Добрый или злой?

Артем: Скорее, запутавшийся. У него же семья такая...

Вика: Да, его жалко немного. Но всё равно гадкий.

Артем: Ну, у него был выбор.

Диалог 7: Джинни Уизли

Настя: Джинни – классная! Боевая девчонка.

Андрей: Да, и в квиддич круто играет.

Настя: Она мне нравится больше всех.

Диалог 8: Дамблдор

Миша: Дамблдор – гений, но он столько всего скрывал!

Юля: Так надо было. Чтобы спасти мир.

Миша: Ну да, логично. Но всё равно... Как-то нечестно.

Юля: Жизнь нечестная.

Диалог 9: Снейп

Рома: Снейп – самый крутой! Такой загадочный.

Даша: Да, он мне сразу понравился.

Рома: Он просто герой.

Диалог 10: Золотое трио

Кирилл: Гарри, Рон, и Гермиона – они как настоящие друзья. Всегда вместе.

Света: Да, это самое главное.

Кирилл: В настоящей дружбе есть настоящая сила.

Диалог 11: Гарри Поттер

Аня: Гарри, конечно, главный герой, но иногда меня раздражает его упрямство.

Боря: Да ладно, без упрямства он бы Волдеморта не победил!

Аня: Это да, но сколько раз он рисковал зря?

Боря: Ну, он же подросток. Кто не ошибается в его возрасте?

Аня: Согласна, но все равно... Иногда хочется ему сказать: "Будь умнее!"

Боря: Зато он очень смелый. Это главное.

Аня: Смелый – да. И верный своим друзьям.

Диалог 12: Рон Уизли

Катя: Рон такой смешной и неловкий. Он мне больше всех нравится.

Дима: Да, Рон крутой. Он самый настоящий из всей троицы.

Катя: И всегда поддержит Гарри, даже когда ему самому страшно.

Дима: А как он боится пауков! Это просто умора!

Катя: Ага! Но все равно он храбрый.

Дима: Точно! Он не идеальный, но это делает его классным.

Диалог 13: Гермиона Грейнджер

Лена: Гермиона – умница. Без нее Гарри бы точно пропал.

Петя: Согласен, она гений. Но иногда она слишком правильная.

Лена: Ну, кто-то же должен следить за порядком!

Петя: Это да, но иногда она душила немного.

Лена: Зато она всем помогла!

Петя: Без вопросов. Гермиона – сила!

Диалог 14: Северус Снейп

Ольга: Снейп – самый загадочный персонаж. Я до конца не понимала, хороший он или плохой.

Слава: Да, он двойной агент! Очень крутой поворот.

Ольга: И как он любил Лили... Очень трагичная история.

Слава: Он был сложным, это точно.

Ольга: И очень сильным волшебником.

Слава: Один из самых.

Диалог 15: Драко Малфой

Соня: Драко мне сначала не нравился, а потом стало жалко.

Миша: Да, он ведь тоже жертва обстоятельств.

Соня: Его семья давила на него.

Миша: У него не было выбора.

Соня: В конце он все равно сделал правильный выбор.

Диалог 16: Дамблдор

Таня: Дамблдор – самый мудрый.

Андрей: Это точно, но он кое-что скрывал...

Таня: Во благо. Он знал, что делал.

Андрей: Надеюсь, он был прав.

Таня: А кто еще?

Диалог 17: Волдеморт

Вика: Волдеморт - такой злой!

Игорь: Он крутой злодей!

Вика: Но я его ненавижу.

Игорь: Он боялся смерти.

Вика: Это его слабость.

Диалог 18: Факультеты Хогвартса

Настя: Я бы хотела быть в Гриффиндоре!

Кирилл: А я в Слизерине! Они крутые!

Настя: Но они злые!

Кирилл: Не все!

Диалог 19: Квиддич

Люба: Квиддич - классный!

Гена: Да, жалко, что в жизни так не поиграть!

Люба: Это было бы опасно.

Гена: Но весело!

Диалог 20: Книги против фильмов

Рита: Книги намного лучше!

Паша: Согласен!

Рита: В фильмах много чего упустили.

Паша: Да, жаль.