ПРАВИТЕЛЬСТВО РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное автономное образовательное учреждение высшего образования

Национальный исследовательский университет «Высшая школа экономики»

Факультет гуманитарных наук

Образовательная программа «Фундаментальная и компьютерная лингвистика»

КУРСОВАЯ РАБОТА

Грамматика ошибок: исследование в контексте нейросетевых моделей обработки языка.

Grammar of Errors: A study in the context of neural network models for natural language processing

Студентка 2 курса
группы №231
Монина Вера Евгеньевна
Научный руководитель:
Выренкова Анастасия
Сергеевна, Доцент

Москва, 2025 г.

Оглавление

1. Введение

В последнее время наблюдается стремительный рост числа нейросетевых моделей, разрабатываемых для обработки естественного языка. Эти модели демонстрируют высокие результаты в решении таких задач, как генерация текста, машинный перевод, суммаризация, извлечение именованных сущностей и других, традиционно относимых к области языкового моделирования.

Для решения подобных задач активно применяются различные архитектуры глубоких нейронных сетей, включая сверточные нейронные сети (CNN), рекуррентные сети с механизмами долгосрочной и краткосрочной памяти (LSTM и GRU), а также модели на основе архитектуры Transformer.

Однако несмотря на высокие показатели производительности, остаётся неясным, насколько поведение таких моделей действительно сопоставимо с человеческим мышлением и языковым восприятием. Именно поэтому проблема интерпретируемости становится всё более актуальной (Belinkov et al., 2023): важно понять, какие знания модели усваивают, как они это делают и насколько процесс “освоения” языка моделями схож с аналогичным процессом у людей. Классический метод оценки эффективности моделей путем сравнивания с правильным ответом уже не эффективен. В связи с этим сейчас вовсю развивается такая область, как пробинг. Исследуется поведение моделей по сравнению с поведением людей в различных психолингвистических экспериментах.

Помимо того, что исследования в этой области позволяют проверить, насколько нюансы человеческого языка воспроизводятся у искусственного интеллекта. Это также позволяет удешевить и ускорить психолингвистические эксперименты, на проведение которых требуется много времени и денег, чтобы провести сами эксперименты и заплатить за прохождение испытуемым. Использование языковых моделей в этих задачах позволяет оперативно получить предварительные оценки лингвистической приемлемости или предсказуемости языковых конструкций, тем самым предоставляя исследователям быстрый доступ к гипотезам о функционировании языка. И, наконец, изучение допускаемых ошибок позволяет уточнить методологию оценки языковых моделей, потому что ошибки маркируют случаи отклонения от нормы и тем самым дают представление о действующих языковых правилах[1].

[1] весь код для экспериментов и обработки данных доступен по ссылке: https://github.com/VeraMonina/kursovaya_2025

2. Роль предсказуемости в психолингвистике

Понимание и распознавание речи обычно происходит инкрементально – человек по мере поступления речевого сигнала пошагово интегрирует каждое новое слово в какое-то формирующееся синтактико-семантическое представление (Lyu B. et al 2024) . При этом важно понимать, что в самом начале высказывания может возникать несколько конкурирующих интерпретаций, особенно в условиях синтаксической или лексической неоднозначности. Такие случаи вызывают когнитивные затруднения и активно изучаются в психолингвистике при помощи экспериментальных методов.

Одним из способов эмпирического исследования того, как слова обрабатываются, в каких контекстах этот процесс может быть легче, как они встраиваются в наше представление, является оценка движения глаз при чтении. При анализе обычно выделяют два типа параметров, которые в наибольшей степени влияют на результат эксперимента: статичные, которые остаются неизменными вне зависимости от контекста, и динамические, зависящие от конкретного контекста. К статичным параметрам относятся, например, длина слова в начальной форме (измеряемая посимвольно) и абсолютная частотность данного слова в языке (вне зависимости от общего контекста). В качестве динамического параметра рассматривается предсказуемость слова — в психолингвистике это вероятность правильного угадывания следующего слова в тексте или предложении на основе предыдущего контекста.

Доказано, что вне зависимости от исследуемого языка, предсказуемость, длина и частотность оказывают наибольшее влияние на то, насколько легко слово будет обрабатываться при чтении. Под «легкостью» понимается короткое время фиксации взгляда на слове и возможность его пропуска при чтении. Также было показано, что слова с высокой предсказуемостью обрабатываются быстрее, чем менее предсказуемые (Ashby et al., 2005; Ehrlich & Rayner, 1981 и др.).

Недавнее исследование (Laurinavichyute, A.K., Sekerina, I.A., Alexeeva, S., 2019) на базе русского языка показало, что предсказуемость, как и в остальных языках, влияет на некоторые аспекты при чтении. Во-первых, с ростом предсказуемости уменьшается вероятность возврата взгляда к предыдущему слову. Во-вторых, вместе с длиной слова предсказуемость влияет на вероятность фиксации взгляда на целевом

слове[2]: чем выше предсказуемость слова и чем оно короче, тем чаще оно будет пропускаться при чтении. В-третьих, наблюдается связь между предсказуемостью слова и временем фиксации взгляда на нем, а также общей длительностью чтения — более предсказуемые слова обрабатываются быстрее.

Предсказуемость слов обычно исследуется при помощи close-процедуры (Taylor, 1953). В рамках этого метода испытуемым предъявляется часть предложения или текста с пропущенным словом, которое они должны предсказать. Как правило, исследуются изолированные предложения, в которых не предсказываются первые и последние слова. В большинстве психолингвистических исследований применяется бинарная система оценки ответов: 1 – если предсказанное слово полностью совпадает с целевым, и 0 – во всех остальных случаях.

Приведем пример того, как может выглядеть психолингвистический эксперимент на предсказуемость. Испытуемому даются незаконченные предложения, как в примерах (1) и (2), которые надо продолжить, в качестве ответа вводится только одно слово.

(1) В качестве примера приводится ___

(2) В вопросе послышался упрёк ___

В типичном психолингвистическом эксперименте правильными ответами для каждого из предложений (см. полные контексты в примерах (3) и (4)) были бы только слова «жирность» и «командиру» соответственно. При этом слова, близкие по значению к целевым — «рецепт» и «генералу», хотя и вполне уместны в данных контекстах, рассматривались бы как неправильные ответы.

(3) В качестве примера приводится <*жирность*>[3] куриного бульона.

(4) В вопросе послышался упрёк <*командиру*>, словно он был виновником происшедшего.

3. Связь предсказуемости и нейросетевых моделей, обзор литературы

Как было обнаружено в последних исследованиях (Wilcox et al., 2020; Merkx and Frank, 2020), чем лучше нейросеть предсказывает следующее слово, тем лучше это отражает человеческое поведение при чтении. В одном из исследований, проведенном

[2] Далее мы будем пользоваться терминами «таргетное слово» или «целевое слово» для описания тех слов, которые надо предсказывать в рамках эксперимента

[3] здесь и далее таргетные слова будут помещаться в <скобки> и выделяться курсивом

на моделях разной архитектуры (n-gram модели, RNN, LSTM, Transformers), рассматривался разный объем обучающего материала. Было показано, что результаты Transformers на данный момент с наибольшей степенью отражают поведение людей.

Также в недавнем исследовании (Oh & Schuler, Findings 2023), проведенном на основе английского языка, было показано, что результаты моделей, которые были обучены на 2 миллиардах токенах и более, соответствуют поведению человека, чем аналогичные нейросетевые модели, но с меньшим количеством параметром. Но при этом, если смотреть на модели, обученные на слишком большом наборе данных (в исследовании в числе рассматриваемых моделей, была исследована модель, обученная на 300 миллиардов токенах) результаты начинают вновь сильно расходиться с результатами людей. Иными словами было показано, что объем обучающего датасета может влиять на корреляцию между поведением моделей и людей, как положительно, так и отрицательно в зависимости от объема параметров.

4. Гипотезы

Гипотеза 1. Предполагается, что такие основные психолингвистические параметры, как длина слова и его частотность, будут влиять на точность предсказаний и людей, и моделей, но это влияние может быть выражено в разной степени. В связи с этим будем рассматривать и то, как длина (см. раздел 8.4) и частотность (см. раздел 8.5) влияет и на точность предсказаний, и на то, как эти же параметры могут в целом влиять на качество предсказаний. Также будет рассмотрен такой параметр, как частотность биграммы, в которую входит предсказание и предыдущее слово в предложении. Предполагается, что этот фактор может влиять на семантическую составляющую успешности предсказаний (раздел 8.5.3).

Гипотеза 2. Ожидается, что увеличение объёма контекста, доступного испытуемому (человеку или модели), может положительно сказываться на качестве предсказаний: чем больше контекста будет видеть «испытуемый» , тем выше качество ответов будет даваться (см. раздел 8.6).

Гипотеза 3. Возможно, у людей будет наблюдаться больше ошибок, связанных с несоответствием предсказанной части речи с ожидаемой, поскольку человеку доступно большее число сложных и комплексных синтаксических конструкций, что расширяет выбор возможных вариантов. В отличие от этого, модели будут давать более прямолинейные и менее разнообразные варианты, что может снизить общую долю

ошибок по сравнению с долей ошибок людей. Также ожидается, что различные части речи обладают разной степенью предсказуемости, независимо от типа испытуемого — будь то человек или языковая модель (см. разделы 8.2 и 8.3).

Гипотеза 4. Эта гипотеза заключается в том, что при предсказании гораздо тяжелее попасть в ожидаемую часть речи целевого слова, нежели чем выбрать лексически подходящий вариант. Иными словами, ошибки, связанные с неверной частеречной принадлежностью предсказаний, могут рассматриваться как более значимые с точки зрения общей сложности предложения. Также будут проанализированы отдельные типы ошибок и те факторы, которые могут повлиять на их появление. Ожидается, что в связи с менее креативным поведением у моделей доля всех ошибок может быть ниже, чем у людей (см. разделы 8.1, 8.7, 8.8 и 8.9).

Гипотеза 5. Последняя гипотеза состоит в том, что BERT покажет более высокие результаты по сравнению и с моделями архитектуры Transformers, и с людьми. Это объясняется тем, что в отличие от классических генеративных трансформеров, BERT использует двунаправленную обработку контекста, что позволяет ему учитывать информацию как слева, так и справа от целевого слова.

5. Данные для исследования

5.1 Предложения, на которых производились эксперименты

Исследование проводилось на 144 предложениях, взятых из Русского корпуса предложений[4], в котором собраны данные о движении глаз взрослых носителей при чтении на русском языке.

Предложения создавались на основе 144 таргетных слов, которые были рандомно выбраны из базы данных StimulStat[5] (Alexeeva, Slioussar, & Chernova, 2017). Отбор слов осуществлялся в соответствии с дизайном эксперимента, разработанным для Potsdam Sentence Corpus[6], включающим три основных лексических параметра: длина слова (число символов в лемме), частотность (абсолютная частотность слова в языке) и часть речи. Все параметры варьировались по следующей схеме: 3 вида длины × 2 вида частотности × 3 части речи. Были взяты слова трех наиболее частотных не служебных частей речи (существительные, глаголы, прилагательные), трех видов длины (short - от

[4] Russian Sentence Corpus
[5] https://stimul.cognitivestudies.ru
[6] Potsdam Commentary Corpus (PCC)

3 до 4 символов, medium - от 5 до 7 символов, long - от 7 символов) и двух видов частотности (low, high). Всего среди целевых слов оказалось 79 существительных, 33 глагола и 32 прилагательных. Неравномерное распределение связано со спецификой русского языка, в котором, практически, невозможно найти короткое прилагательное или глагол (от 3 от 4 символов), в связи с этим было увеличено количество существительных. По длине и частотности распределение таргетов равномерное (48 × 48 × 48). Далее эти слова служили стимулами для составления предложений. Также именно эти стимулы предсказывались в экспериментах.

Далее были выбраны предложения из Национального корпуса русского языка[7] (НКРЯ), в которых есть один из стимулов (при этом учитывалось, что стимул не может занимать первую или последнюю позицию в предложении). При выборе предложений также учитывалась их длина, чтобы они были не слишком длинными или короткими. Поэтому некоторые предложения, могли быть специально укорочены и упрощены (см. примеры (5a) и (5b), где сначала идет оригинальное предложение, а потом укороченное). Затем подобранные предложения были оценены контрольной группой, состоящей из 750 человек, на приемлемость и подкорректированы относительно результатов их оценки. В итоге, было получено 144 предложения (все предложения можно посмотреть в приложении 4), в каждом из котором есть по одному целевому слову, про которое известны его длина, частотность и часть речи. Стоит отметить, что в данном исследовании слово «который» рассматривается как прилагательное.

(5) a. В болотах млел еще желтый кислый <*лед*>, но на берегах уже появилась из-под снега прошлогодняя трава и груды торфа. [Юрий Коваль. У Кривой сосны (1979)]

b. На болотах оставался ещё <*лед*>, но на берегах реки появилась трава.

5.2 Данные о предсказаниях людей

В дальнейшем в нашем исследовании будут использованы данные, собранные другой группой исследователей в рамках работы с вышеупомянутым корпусом русских предложений. Эти данные получены в исследовании Laurinavichyute, A.K., Sekerina, I.A., Alexeeva, S. (2019), где помимо предсказуемости слов также фиксировалось движение глаз при чтении. Эксперимент, проведённый авторами , несколько отличался

[7] https://ruscorpora.ru/

от нашего: участникам необходимо было предсказывать не только целевые слова, но и все остальные слова в предложении. Опрос проводился онлайн без ограничения на количество предсказанных слов одним человеком. В анализ были включены все анкеты, в которых участники предсказали 20 или более слов. В результате количество предсказаний для разных слов одного предложения может сильно варьироваться. Всего было собрано чуть более 65 тысяч предсказаний.

Для данного исследования были взяты только те предсказания, в которых людям предсказывали целевые слова. После проведенной сортировки осталось 6900 предсказаний. Для каждого предсказания уже было размечено, является ли оно точным попаданием или нет. Далее для каждого предсказания были автоматически размечены его длина, часть речи, совпадает ли части речи данного слова и таргетного (в формате 0/1). Частеречная разметка была проведена при помощи Mystem[8]. Все предсказания были размечены по специальным семантической и грамматической разметкам (см. Раздел 7). Также для каждого ответа была автоматически подсчитана его относительная позиция в предложении (насколько далеко от начала оно находится).

6. Методология

6.1 Эксперимент на моделях с левым контекстом

Основной эксперимент был проведён на трёх моделях: ruGPT-3.5[9], Llama-3.1[10], Qwen/QwQ[11]. Эти модели были выбраны по двум причинам. Во-первых, на сегодняшний день они являются одними из самых мощных и передовых среди больших языковых моделей, находящихся в открытом доступе. Во-вторых, все версии этих моделей обучены на больших объёмах данных и содержат более 2 миллиардов параметров – порог, который в исследовании Oh Schuler (Findings, 2023) был признан критическим, чтобы результаты моделей и людей стали похожими.

Все три модели имеют архитектуру трансформеров, которая сейчас занимает доминирующие позиции в задачах обработки естественного языка (NLP). Эта архитектура отличается от других, которые были основаны на основе рекуррентных вычислений (GRU и LSTM). Преимущество трансформеров заключается в применении

[8] pymystem3PyPIhttps://pypi.org › project › p...
[9] ai-forever/ruGPT-3.5-13B · Hugging Face
[10] unsloth/Llama-3.1-8B-Instruct-GGUF · Hugging Face
[11] Qwen/QwQ-32B · Hugging Face

механизма внимания, который позволяет модели оценивать значимость связи между каждой парой слов в обрабатываемом тексте вне зависимости от его длины.

Дизайн эксперимента полностью повторяет дизайн стандартного психолингвистического эксперимента на предсказуемость. Использовались изолированные предложения, в которых были заранее выбраны исследуемые слова. Моделям подавался левый контекст без учета таргетного слова, и требовалось продолжить предложение. В качестве результата бралось первое слово из предсказанных.

ruGPT-3.5 — это нейросетевая языковая модель для русского языка с 13 миллиардами параметров. Модель обучена на большом корпусе русскоязычных текстов. Llama-3.1 — мультиязычная авторегрессионная модель архитектуры Transformers с 8 миллиардами параметров, длина поддерживаемого контекста - до 131072 токенов. Qwen/QwQ - также модель архитектуры Transformers, состоящая из 64 слоев, с 32.5 миллиардами параметров.

Чтобы получить разнообразные варианты предсказаний от моделей, изменялись четыре основных параметра (см. Приложение 1). Первый параметр - top_k, который изменяется для того, чтобы повысить качество выводимых предсказаний. При выборе следующего слова модель отфильтровывает наименее вероятные варианты предсказаний и выбирает конечный ответ из наиболее вероятных вариантов. То есть, если значение параметра k = 30, то модели будет рассматривать тридцать наиболее вероятных слов. Второй метод сэмплирования - top_p, который также ограничивает множество слов, из которых модель выбирает следующее слово. При использовании данного параметра будут учитываться только те токены, совокупная вероятность которых превышает заданное значение p, все остальные токены отбрасываются и не используются при выборе ответа. Например, при p = 0.95 будут рассматриваться только те токены, суммарная вероятность которых равна или превышает 0.95. Третий параметр, который изменялся - temperature, он отвечает за креативность предсказаний, тем самым делая ответы более разнообразными. Чем выше температура (то есть, чем ближе значение данного параметра к 1), тем более разнообразными будут ответы. И последний метод, который применялся в исследовании - beam search. При применении этого алгоритма выбирается заданное количество различных путей, которые будут рассматриваться на каждом шаге генерации. Без применения данного алгоритма будет

работать «жадная» генерация текста, то есть, модель будет выбирать следующий токен, у которого вероятность появления после предыдущего максимальная.

От каждой модели было получено от 1 до 3 предсказаний для каждого предложения. В данной работе все предсказания будут проанализированы вместе, без деления на отдельные модели. Результаты, полученные для каждой модели по отдельности, можно посмотреть в Приложении 2.

Всего было получено 1008 предсказаний, среди которых около 100 предсказаний оказались невалидными, потому что модели не смогли предсказать существующее слово на место таргета.

Для каждого предсказания были автоматически размечены те же параметры, что и для предсказаний людей (см. раздел 5.2). В дополнение также было размечено наличие точного попадания (тоже в формате 0/1), чтобы облегчить дальнейшую семантическую разметку.

6.2 Эксперимент на модели с правым и левым контекстом

Второй дополнительный эксперимент проводился на двух версиях модели RuModernBERT: RuModernBERT-small[12] и RuModernBERT-base[13].

RuModernBERT – это последняя вариация модели RuBERT, опубликованная 19 февраля 2025 года, в двух версиях: с 35 миллионами и 150 миллионами параметров. Обе версии поддерживают контекст длиной до 8 тысяч токенов и обучены на корпусе в 2 триллиона токенов. Для них был также разработан новый токенайзер и по сравнению с предыдущими версиями увеличена длина контекста для улучшения качества понимания и генерации текста.

Отметим, что предварительная тренировка BERT-а проходит на неразмеченных данных и состоит их двух этапов. Один из них заключается в создании «маскированной языковой модели»: 15% входных слов заменяются служебным токеном [MASK], и модель обучается восстанавливать эти скрытые слова. Именно этот алгоритм позволяет моделям этой архитектуры эффективно заполнять пропуски в предложениях и текстах.

В рамках эксперимента использовались предложения, описанные в разделе 5.1, в которых целевое слово было заменено специальным токеном [MASK]. Модели ставилась задача предсказать слово, скрытое за этим токеном.

[12] deepvk/RuModernBERT-small · Hugging Face
[13] deepvk/RuModernBERT-base · Hugging Face

Главное отличие данного эксперимента от предыдущего заключается в том, что модели имели «видели» как к левый, так и правый контекст, что значительно улучшает точность и качество предсказаний. Отметим, что хотя в психолингвистических исследованиях предсказуемости иногда проводят эксперименты, в которых участники видят обе части предложения или текста, такие подходы встречаются значительно реже, чем стандартные, которые были описаны в предыдущих разделах.

Всего было получено 288 предсказаний, из которых 10 оказались невалидными. Все остальные предсказания были размечены и проанализированы по тем же критериям, что предсказания и в предыдущем эксперименте.

7. Разметка полученных результатов

В нашем исследовании мы отошли от принятого в психолингвистике анализа, в котором все предсказания делятся на те, которые полностью совпадают с таргетом, и все остальные, потому что предполагается, что для одного контекста может быть дано несколько разных предсказаний, которые одинаково корректны (см. примеры (1) и (2)). Эту вариативность можно попытаться отразить при помощи более дробной семантической разметки каждого предсказания, тем самым позволяя учитывать частично корректные или альтернативные, но всё ещё уместные ответы, которые ранее были бы рассмотрены как «неверные».

Помимо более дробной семантической разметки, была введена отдельная грамматическая разметка, для того, чтобы эти две области ошибок никак не пересекались, потому что можно, например, предсказать неприемлемое семантическое слово, но оно будет верно согласовано по роду, или же можно предсказать приемлемое слово, которое при этом будет стоять в неверном падеже. Чтобы не потерять за очень узкой и сжатой разметкой все такие нюансы, семантика и грамматика были разведены на две непересекающиеся разметки.

Часть тегов, как из семантической разметки, так и из грамматической, была взята из классификации ошибок в Russian Learner Corpus[14].

[14] http://www.web-corpora.net/RLC/help

7.1 Семантическая разметка

В семантической разметке основное внимание уделяется взаимной сочетаемости слов, приемлемости коллокаций. Одному предсказанию может соответствовать только один из тег (см. Таблицу 1).

Таблица 1. Система семантических тегов.

Тег	Описание	Предложение	Предсказание
Exact	Тег ставится только в том случае, если предсказанное слово полностью совпадает с таргетным.	У моего отца был счёт в швейцарском <*банке*>, он был лесоторговец!	*банке*
Close	Ставится на предсказание в том случае, если у таргетного слова и предсказанного один корень, но формы слова различаются, при этом предсказанное слово должно подходить под контекст предложения. Например,этим тегом могут размечаться видовые пары одного глагола.	Телята быстро <*росли*>, превращаясь в ласковых коров.	*растут*
Accept	Тег описывает те ситуации, когда предсказанное слово не совпадает с таргетным, но при этом оно могло бы быть употреблено в данном контексте. Важно, что данный тег также описывает те редкие случаи, когда часть речи таргетного слова и предсказанного не совпали, но такое употребление приемлемо.	В сюжете этого фильма какие-то <*осы*> устраивают себе гнёзда в дереве.	*птицы*
		Старый шкаф <*орехового*> дерева был явно не отсюда	*из*
Lex	Тег ставится, если часть речи таргетного слова и предсказанного совпали, но предсказанное слово нельзя употребить в соответствующем контексте.	Ему удалось вскрыть банку об острый край <*бампера*> своего автомобиля.	*стола*
POS_Error	Тег ставится во всех случаях, когда части речи целевого слова и предсказанного не совпадают, и при этом предсказанное нельзя считать допустимым вариантом (Accept).	На вторичном рынке жилья <*розетки*>	*наблюдается*

		заклеивают обоями во время ремонта.	
Error	Тег описывает все предсказания, не являющиеся существующими словами.	Когда родители *<пригрозили>* не взять ее с собой, Маша очень расстроилась.	*ар*

7.2 Грамматическая разметка

При разметке данного класса ошибок (см. Таблица 2) встречались предсказания, в которых было допущено больше одной ошибки, в таком случае ставилось несколько тегов, описывающих каждую из ошибок.

Стоит отметить, что все ошибки можно поделить на четыре большие группы. В первую группу можно включить все предсказания, которые стоят в абсолютно верной грамматической форме. Ко второй группе можно отнести предсказания, которые верны относительно левого контекста (видимого в эксперименте), но при учете правого контекста становятся неграмматичными (например, при учете дополнительной информации становится понятно, что требуется другая числовая форма). К третьей группе можно отнести предсказания, в которых допущены ошибки при условии, что вся необходимая информация (например, вершина предложной группы) содержится в левом контексте. То есть испытуемый совершил «реальную» ошибку, а не потому что ему в рамках эксперимента не была дана какая-то необходимая синтаксическая информация. И к последнему классу можно отнести ошибки, когда испытуемый не смог попасть в часть речи целевого слова, потому что чаще всего такие предсказания ведут к нарушению синтаксического строя предложения.

Таблица 2. Система грамматических тегов.

Тег	Описание	Предложение	Предсказание	Комментарий (при наличии)
NO_Error	Тег ставится при отсутствии каких-либо грамматических ошибок.	Вспоминая *<журчание>* водяных струй, мы непременно	*слон* (Семантический тег: Accept)	

		подумаем о фонтанах.		
Syntax	Тег практически всегда ставится вместе с семантическим тегом POS_Error, отражая нарушение какого-либо синтаксического строя, предполагаемого в изначальном предложении.	Какие главные лекарства должны *<входить>* в аптечку автомобилиста?	*всегда* (Семантический тег: POS_Error)	
		Какие главные лекарства должны <входить> в аптечку автомобилиста?	*быть* (Семантический тег: Accept)	Предсказания такого типа размечаются этим грамматическим тегом, потому что глагол *быть* управляет предложной группой в предложном падеже, а не в аккузативе, поэтому не смотря на то, что семантически этот глагол мог бы использовать в этом контексте, у него стоит грамматический тег Syntax.
Gender	Данный тег применяется к предсказаниям, которые не соответствуют грамматическому роду в контексте предложения. Тег ставится, если предсказанное слово —	У мамы есть *<подруга>*, которая живет прямо напротив здания театра.	*сын* (Семантический тег: Accept)	Данный пример интересен тем, что предсказанное слово попадает в правильную семантическую область целевого

	существительное (для ошибок в согласовании про роду есть тег AgrGender).			слова (в первую очередь важна одушевленность предсказания), однако грамматически оно неверно. Если заменить предсказанное слово на ближайшее по смыслу, но женского рода (дочь), то грамматическая ошибка исчезает.
Num	Тег описывает предсказанные слова, стоящие в неверной числовой форме (несоответствующей контексту), может ставиться только если предсказанное слово — существительное.	Музыканты играли на похоронах, разгружали <*вагоны*>, жили бедно.	*гробы* (Семантический тег: Accept)	
Gov	Ставится при наличии ошибки в глагольном управлении, может ставиться только если предсказанное слово — существительное.	На болотах оставался еще <*лед*>, но на берегах реки появилась трава.	*тоннеля* (Семантический тег: Lex)	
AgrGender	Ставится на предсказания, если они неверно согласованы по роду (например, прилагательные,	У нас в Волгограде многие придерживаются	*такого* (Семантический тег: Accept)	

	причастия или глаголы в прошедшем времени).	<*иной*> точки зрения.		
AgrNum	Ставится на предсказания, если они неверно согласованы по числу (например, прилагательные, глаголы).	Покуда я нахожусь у власти, я буду предметом <*ярых*> нападок соперников.	*пристального* (Семантический тег: Lex)	
AgrCase	Ставится при наличии ошибки в согласовании по падежу (на прилагательных, местоимениях и причастиях).	Мне нравится сын коллеги, <*который*> недавно заходил в наш отдел.	*которого* (Exact)	
		Мне нравится сын коллеги, <*который*> недавно заходил в наш отдел.	*которую* (+ AgrGender, семантический тег: Exact)	
Translit	Тег используется, если предсказанное слово записано латиницей.	Когда она в самолете <*летела*> домой, читать не было сил.	*uvidela* (Семантический тег: Lex)	
Misspell	Тег описывает все орфографические ошибки.	Причиной аварии был мобильный <*телефон*>, который отвлекал водителя от дороги.	*телеффон* (Семантический тег: Exact)	

Brev	Тег описывает ошибки в употребление кратких или полных прилагательных (то есть ситуации, когда было употреблено краткое прилагательное вместо полного, либо полное вместо краткого)	Он был очень <*неопытным*> дипломатом и большим мечтателем.	*рад* (Семантический тег: Lex)	
Asp	Ошибка заключается в выборе неверной видовой форме глаголы.	Елена сидела в кресле, молодая Мурка урчала у нее на коленях.	прыгнула (Семантический тег: Accept)	
Tense	Ошибка во временной форме глагола (например, когда по правому контексту ясно, что требуется прошедшее время, а предсказанное слово стоит в настоящем).	Когда она в самолете <летела> домой, читать не было сил.	летит (Семантический тег: Close)	
Error	Тег описывает все предсказания, которые не являются существующими словами (тег повторяется с тегом из семантической разметки).	Когда родители <*пригрозили*> не взять ее с собой, Маша очень расстроилась.	*ap* (Семантический тег: Error)	

8. Анализ

8.1 Общий анализ

Рассмотрим общие результаты (см. рисунки 1 и 2), полученные для людей и моделей. На этих графиках представлены доли, показывающие, как были распределены все предсказания по семантическим и грамматическим тегам относительно всех данных предсказаний. Иными словами, каждая доля отражает, как часто среди результатов встретились предсказаний с конкретным тегом, что позволяет сравнить

качество предсказаний людей и моделей. Отдельно рассматривается семантическая предсказуемость, отдельно грамматическая.

Можно отметить, что больше 50% предсказаний грамматически корректны, как у людей, так и у моделей. Также и моделями, и людьми было дано практически в два раза больше предсказаний с семантическим тегом Accept, чем с тегами Close и Exact, то есть приемлемых вариантов было дано гораздо больше, чем точных попаданий.

Рисунок 1. Общие результаты по семантической приемлемости .

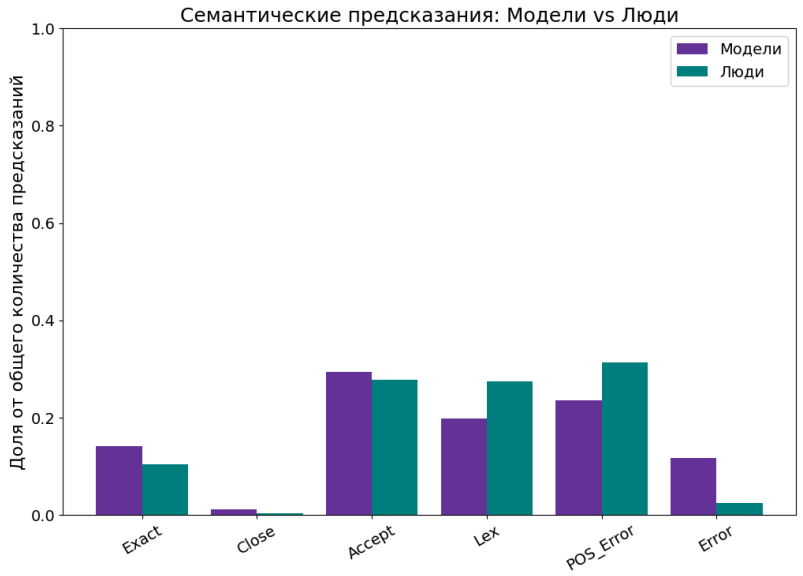
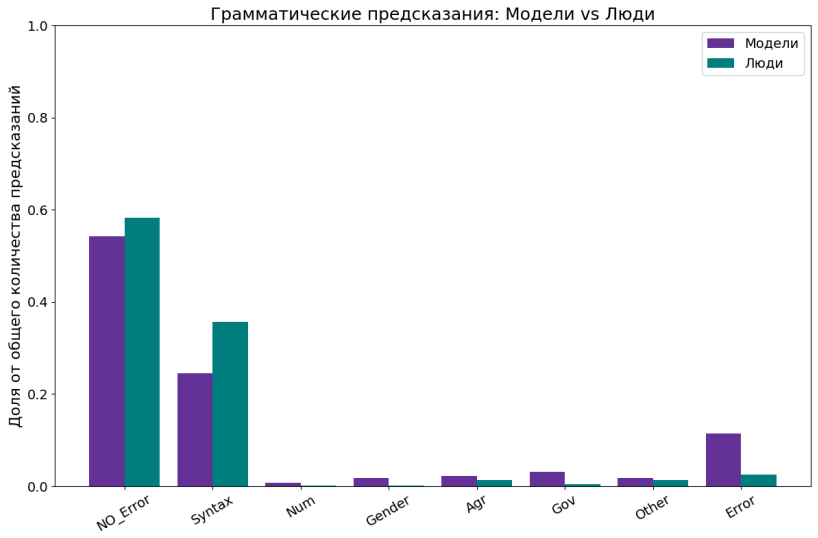


Рисунок 2. Общие результаты по грамматической корректности.



Анализируя предсказания с точки зрения семантической приемлемости (рис. 1), стоит заметить, что у моделей семантически приемлемых предсказаний больше, чем лексически неподходящих (Lex) и предсказаний с ошибками, заключающимися в выборе неподходящей части речи (POS_Error), которые в свою очередь ведут к синтаксическим ошибкам (Syntax). В то время как у людей доля лексических и ошибок

в части речи больше, чем у моделей. При этом у людей доля ошибок, когда было предсказано несуществующее слово (Error) сильно ниже, чем у моделей.

Обращаясь к рисунку (2), можно заметить, что самая распространенная грамматическая ошибка, не считая нарушения синтаксического строя - это ошибка в управлении (Gov). Все ошибки, допущенные в согласовании (Agr)[15], на графике были объедены в один тип ошибок, так как каждый отдельный тип ошибки был представлен, скорее, единичными примерами. Остальные ошибки, как, например Brev, Asp и др., были отнесены в столбец «other». Как мы видим, таких ошибок было допущено, практически, столько же, сколько и ошибок типа Agr, Gender (родовая принадлежность слова не соответствует контексту) и Num (числовая форма слова не соответствует контексту). Более высокая доля ошибок типа Syntax у людей объясняется тем, что у них было сделано больше ошибок в определении ожидаемой части речи. Также можно отметить, что у людей доля предсказаний без грамматических ошибок немного выше, чем у моделей. Также процент ошибок, типа Gov, Agr, Num и Gender, ниже, чем у моделей, что указывает на более качественные грамматические ответы.

8.1.1 Вывод из раздела 8.1

Из этого раздела можно сделать вывод, что **количественная оценка ответов, данных людьми ниже, чем оценка ответов, сгенерированных моделями, что подтверждает гипотезу 4**. Поэтому в следующих разделах будет проводиться не только количественный анализ, но и качественный.

8.2. Анализ предсказаний в зависимости от части речи таргетного слова

В таблице 3 представлены результаты анализа предсказуемости целевых частей речи – существительных, глаголов и прилагательных – для моделей и участников эксперимента. В столбце «correct pos» отражен процент предсказаний, у которых часть речи совпала с частью речи таргетного слова (например, для всех предложений с целевыми существительными указан процент предсказаний, где модель или человек также предсказали существительное вне зависимости от семантической и грамматической правильности этих предсказаний). Аналогичные расчеты были произведены для прилагательных и глаголов.

В столбце «correct sem» приведён процент предсказаний, которые не только совпали по части речи с целевым словом, но и семантически корректны.

[15] AgrNum, AgrGender, AgrPers

Семантическая корректность оценивалась на основе разметки категорий Exact, Close и Accept, которые были объединены вместе. Аналогичным образом в столбце «correct gram» представлен процент предсказаний, совпавших по части речи и, которые грамматически корректны (то есть для которых стоит тег NO_Error).

Таблица 3. Предсказуемость частей речи.

	Модели			Люди		
	correct POS	correct sem	correct gram	correct POS	correct sem	correct gram
noun	79,84	55,93	68,38	71,19	43,05	69,92
verb	63,32	38,69	55,28	79,24	36,73	54,07
adj	42,02	27,66	30,85	26,72	18,54	18,54

Следует отметить, что в данном анализе не рассматриваются случаи, когда предсказания семантически приемлемы, но не совпадают с частью речи целевого слова (например, предсказание наречия вместо прилагательного). Такие примеры будут рассмотрены далее, в разделе 6.3.

Как видно из таблицы (3), грамматические характеристики предсказываются лучше, чем семантические, что свидетельствует о том, что модели и люди более уверенно справляются с выбором грамматических форм, нежели чем с выбором контекстуально подходящих лексем.

Данные, представленные в таблице, свидетельствуют о том, что среди анализируемых частей речи наивысшие показатели совпадения части речи между предсказанным и целевым словом демонстрируют существительные. Более того, именно для существительных зафиксированы наиболее высокие значения как семантической, так и грамматической корректности. Глаголы уступают существительным по всем трём метрикам (процент верных попаданий в часть речи, в семантику, в грамматику). Прилагательные демонстрируют наихудшие процентные соотношения из всех трех частей речи.

Примечательно, что для прилагательных наблюдается минимальная разница между долями грамматически и семантически корректных предсказаний – около трёх процентов. Это указывает на то, что хотя предсказания на месте прилагательных достаточно редко совпадают с целевой частью речи (всего в 42,02% случаев), но если модель всё же правильно идентифицирует часть речи, то вероятность того, что предсказание будет одновременно семантически и грамматически приемлемым,

становится относительно высокой и сбалансированной. Иными словами, ошибки в прилагательных носят обладают такой особенностью : очень сложно попасть в нужную часть речи, но если угадать удалось, то предсказание будет с большей вероятностью грамматически и семантически правильным.

Переходя к анализу предсказаний людей, можно заметить, что общая тенденция соотношения между семантически верными и грамматически верными предсказаниями сохраняется. Кроме того, те же тенденции наблюдаются и относительно предсказаний прилагательных с тем отличием, что небольшая разница в предсказуемости семантических и грамматических признаков становится еще более заметной — несмотря на более низкие показатели предсказуемости самой категории прилагательного, доля приемлемых предсказаний и предсказаний без грамматических ошибок совпадает (это не означает, что во всех семантически верных предсказаниях не было допущено грамматических ошибок; это также не означает, что во все грамматически верные предсказания приемлемы).

Несмотря на то, что, на первый взгляд, показатели семантической точности предсказания и точности угадывания ожидаемой части речи людьми кажутся ниже, чем у моделей, при более детальном анализе выявляется важная закономерность. В частности, для существительных, если человек смог верно предсказать часть речи, доля грамматических ошибок среди этих предсказаний существенно ниже, чем у моделей: у моделей разница между долей попадания в целевую часть речи и долей грамматических ошибок составляет около 10%, тогда как у людей – около 3%.Это указывает на то, что человеческий анализ в рамках уже правильно идентифицированной частеречной категории более точен с точки зрения грамматической корректности.

Данная тенденция не сохраняется для глаголов: у людей соотношение грамматических ошибок и предсказаний, попавших в нужную часть речи, увеличивается по сравнению с моделями, но при этом сама категория глагола угадывается чуть лучше.

В целом, результаты людей характеризуются более низкими значениями по точности определения частей речи и их семантической корректности по сравнению с моделями. Это может быть связано с большей синтаксической и лексической вариативностью, которой располагает человек при формировании предсказаний. Рассмотрим следующий пример (6) и предсказания к нему:

(6) Её сын Гриша умер <*бездетным*>, и младший сын остался единственным наследником.

Модели предсказывают слова «рано», «недавно», «сегодня». То есть какие-то односоставные, простые формы, у которых нет никаких зависимых, и которые семантически подходят под контекст. Это очень частотная ситуация, которая указывает на то, что модели часто склонны к выбору более частотных и однозначных вариантов, например, таких как какие-то наречия с временным значением, которые семантически подходят под контекст и имеют высокую вероятность при предсказании. В то же время люди, обладая более широким выбором при предсказании, склонны предлагать более сложные и развернутые синтаксические конструкции — например, «прошлой (весной/летом/зимой)[16]», «месяц (назад)», «от (какой-то болезни)». Такие варианты отражают не только конкретное время, но и причинно-следственные связи. Если бы «окно» рассмотрения предсказаний было шире, чем одно слово, то конструкции такого рода, наверняка попали бы под приемлемые предсказания, но так как в эксперименте рассматривается предсказание только одного слова, засчитать такие предсказания, как что-то подходящее, нельзя.

Рассмотрим еще один пример (7), который также иллюстрирует замечание, приведенное выше:

(7) Один футболист, который получил <*растяжение*>, не участвовал в игре.

Моделями опять предсказываются достаточно очевидные и наиболее вероятные варианты по типу «травму», «растяжение» и «награду», в то время как люди помимо наиболее вероятных вариантов также предсказывают такие варианты, как «красную (карточку)», «желтую (карточку)», «кубок», «штрафной».

То есть в обоих случаях мы видим, что человек располагал несколько достаточно частотных словосочетаний, которые могли бы быть использованы в предложенном контексте, но так как в рамках эксперимента рассматривается только одно слово, все такие предсказания не могут быть засчитаны и чаще всего ведут к нарушению синтаксического строя предложения, что ведет к более низким количественным результатам. В то время как модели стремятся к наиболее вероятному и простому варианту, что повышает их показатели точности угадывания части речи и общие показатели подходящих предсказаний, но надо держать в голове, что эти предсказания более очевидные и менее комплексные, чем у людей.

[16] В скобках указано то, что люди скорее всего ожидали увидеть далее по контексту

При этом при условии правильного распознавания таргетной части речи, люди демонстрируют более высокий уровень грамматической точности, что указывает на более глубокое понимание грамматических правил, задаваемых конкретным контекстом.

8.2.1 Вывод из раздела 8.2

В данном разделе было показано, что **модели лучше предсказывают ожидаемую часть речи**, чем люди. Но в **рамках правильно выбранной части речи люди в среднем допускают меньшее число грамматических ошибок** (см. таблицу (4)). И модели, и люди **лучше всего предсказывают существительные**. Люди предсказывают хуже глаголы (как грамматические признаки, так и семантические), а модели хуже справляются с прилагательными (тоже по обоим параметрам). Также были показаны несколько показательных примеров ((6) и (7)), которые демонстрируют, что **люди дают более креативные и синтаксически сложные предсказания**, чем модели, что подтверждает **гипотезу 3**.

Таблица 4. Процент ошибок при верно угаданной части речи

	Модели		Люди	
	разница между correct POS и correct sem (в %)	разница между correct POS и correct gram (в %)	разница между correct POS и correct sem (в %)	разница между correct POS и correct gram (в %)
noun	23,91	**11,46**	28,14	**1,27**
verb	24,63	8,04	42,51	25,17
adj	14,36	**11,17**	8,18	**8,18**

8.3 Анализ предсказаний альтернативной части речи

Как уже было несколько раз замечено в предыдущих разделах, приемлемыми могут быть не только варианты, которые верно попали в часть речи целевого слова, но и предсказания, не попавшие в ожидаемую часть речи, но которые семантически подходят под контекст. Например, как предсказание к предложению (8):

(8) Существует легенда, что <*Ноев*> ковчег вынесло на вершину этой горы.

Предсказание: *давным-давно*

Как можно увидеть на рисунках (3) и (4), существительные и прилагательные - это те части речи, при предсказании которых допустима небольшая вариативность.

Обычно, такая ситуация возможна, когда вместо прилагательного предсказывается наречие, как в примере (8), а вместо существительного - местоимение (если по контексту нужен одушевленных участник), то есть при замене одной части речи другой, но близкой по возможным выполняемым функциям. В то время, как глаголы не обладают такой вариативностью. Скорее всего это связано, во-первых, с их большей обязательностью в предложении. Во-вторых, в русском языке практически нет других частей речи, которые могли бы выполнять предикативную функцию в полной мере, как это делают глаголы. Поэтому глаголы почти никогда не могут быть грамматически корректно заменены на элементы другой части речи, что и демонстрируют графики.

Рисунок 3. Соотношение приемлемых предсказаний моделей в зависимости от попадания в ожидаемую часть речи.

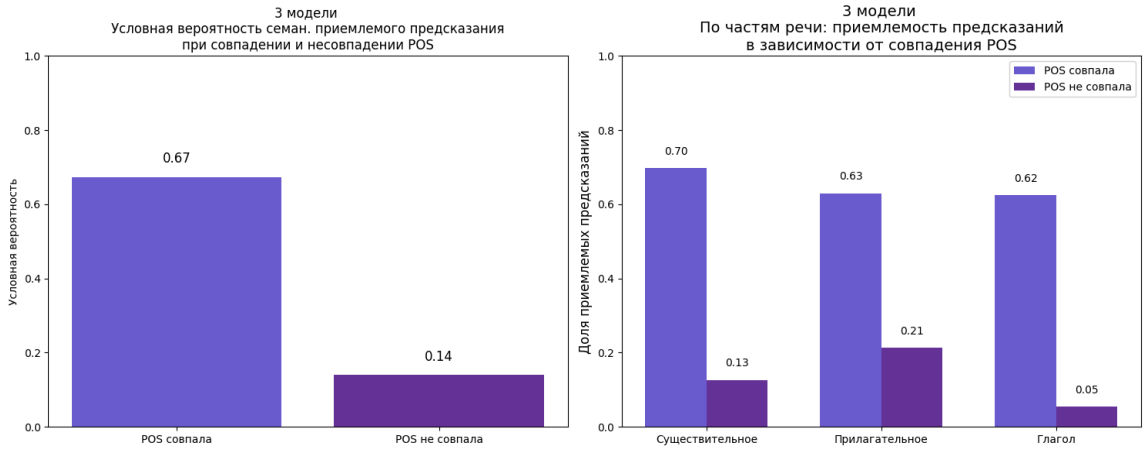
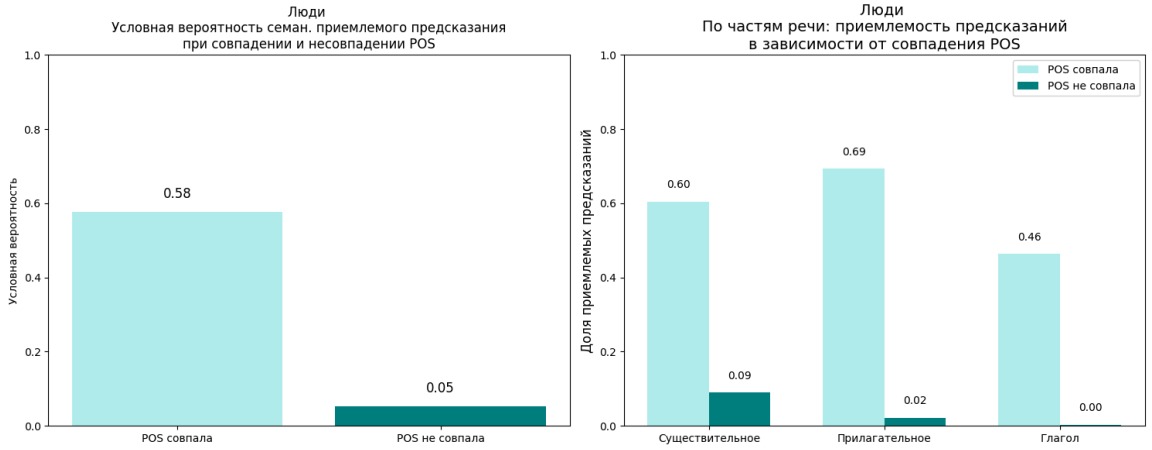


Рисунок 4. Соотношение приемлемых предсказаний людей в зависимости от попадания в ожидаемую часть речи.



Бóльшая доля приемлемых предсказаний, не попавших в ожидаемую часть речи, у моделей объясняется сравнительно большим количеством предсказаний, когда вместо прилагательного было предсказано наречие. Это опять-таки в некоторой степени указывает на то, что их предсказания более простые. То есть, обычно, нет попытки предсказать какое-то словосочетание, в котором сначала идет зависимое, а только

потом вершина (см. предсказания к примерам (6) и (7)), это предсказания – самостоятельные части речи, которые не требует больше ничего и могут самостоятельно функционировать в предложении.

Стоит отметить, что среди предсказаний другой части речи, очень распространена ситуация, когда и люди, и модели предсказывают предлог (вне зависимости от того, какая часть речи была скрыта). В таком случае - это сразу ведет к синтаксической ошибке, потому что зависимого предложной группы нет (см. пример 9).

(9) В числе возможных кандидатов <*называют*> депутата от партии правых.

Предсказания: *на, в*

Видимо, предполагались такие триграммы: «кандидатов на (должность/премию и т.д)», «кандидатов в (президенты/кандидаты и т.д)».

8.3.1 Вывод из раздела 8.3

Раздел 8.3 иллюстрирует, что **при предсказании существительных и прилагательных допустима небольшая вариативность в выборе другой части речи**, которая приводит к приемлемому варианту.

8.4 Влияние длины целевого слова на предсказания

8.4.1 Влияние абсолютной длины слова на приемлемость предсказания

Проверка влияния абсолютной длины таргетного слова (количество знаков в лемме) на семантическую приемлемость предсказания проводилась с помощью подсчета коэффициента Пирсона. Для моделей коэффициент равен -0.025, а p-value = 0.76, что говорит о том, что результаты статистически незначимы и корреляция практически, отсутствует (совсем небольшая и имеет отрицательное значение). Для людей результаты подсчетов равны 0.041 и 0.63, соответственно. Это вновь говорит об отсутствии статистической зависимости и отсутствии явной линейной зависимости. Поэтому в следующем разделе было рассмотрено влияние длины (уже не абсолютной, а поделенной на три категории long, medium и short) не на вероятность приемлемого предсказания, а на вероятность того, что предсказания окажется точным.

8.4.2 Влияние длины слова (long / medium / short) на точность предсказания

На рисунках 5 и 6 представлены результаты проверки гипотезы о том, что короткие слова (от 3 до 4 символов) предсказываются точнее, чем более длинные. Для этого были взяты все семантически корректные предсказания. Далее для слов каждой

длины (long / medium / short) была проанализирована доля точных предсказаний относительно всех приемлемых.

Рисунок 5. Зависимость точности предсказаний людей в зависимости от длины таргетного слова.

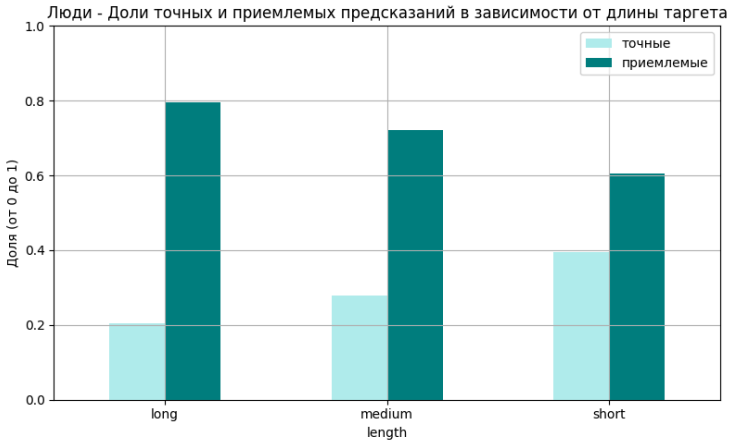
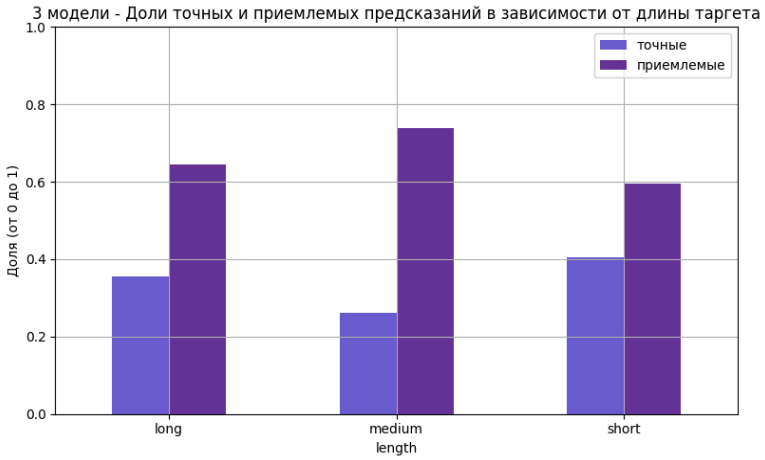


Рисунок 6. Зависимость точности предсказаний моделей в зависимости от длины таргетного слова.



Как для людей, так и для моделей справедливо утверждение, что среди семантически подходящих предсказаний наибольшая доля точных попаданий (отмеченных тегом Exact) приходится на короткие слова — примерно 0.4.

В ответах людей наблюдается явная зависимость точности от длины слова: чем длиннее слово, тем больше доля семантически приемлемых предсказаний (около 0.8 для длинных слов, 0.7 для слов средней длины и 0.6 для коротких).

В ответах моделей такой четкой зависимости выявить не удалось. Наименьшая доля точных предсказаний приходится на таргетные слова средней длины, за ними следуют длинные слова с долей немного ниже 0.4, а затем короткие слова с долей около 0.4.

Остается не вполне ясным, почему у моделей среди слов средней длины наименьшая доля точных предсказаний. Можно было бы предположить, что в эту

категорию попало наибольшее количество прилагательных, но эта теория не подтверждается проверкой. Как среди коротких, так и среди слов средней длины, встречается больше целевых слов – существительных, чем таргетов других частей речи. А как уже было показано, существительные, с точки зрения семантики, обладают наилучшей предсказуемостью по сравнению с другими частями речи.

Также стоит заметить, что люди в целом более чувствительны к длине слова, чем модели. Разница между количеством точных предсказаний людей при long и short таргетных словах практически 20%, в то время как разница между medium и short у моделей – примерно 15% (между long и short еще меньше). То есть у моделей результаты немного сбалансированнее, чем у людей.

8.4.3 Выводы из раздела 8.4

В разделе 8.4 было показано, что явной **корреляции между длиной таргетного слова и успешностью предсказания нет** (см. раздел 8.4.1) как для людей, так и для моделей. Длина таргетного слова влияет только на точность предсказания, то есть чем слово короче, тем с большей вероятностью и модель, и человек сделает предсказание, которое полностью совпадет с таргетным (см. раздел 8.4.2).

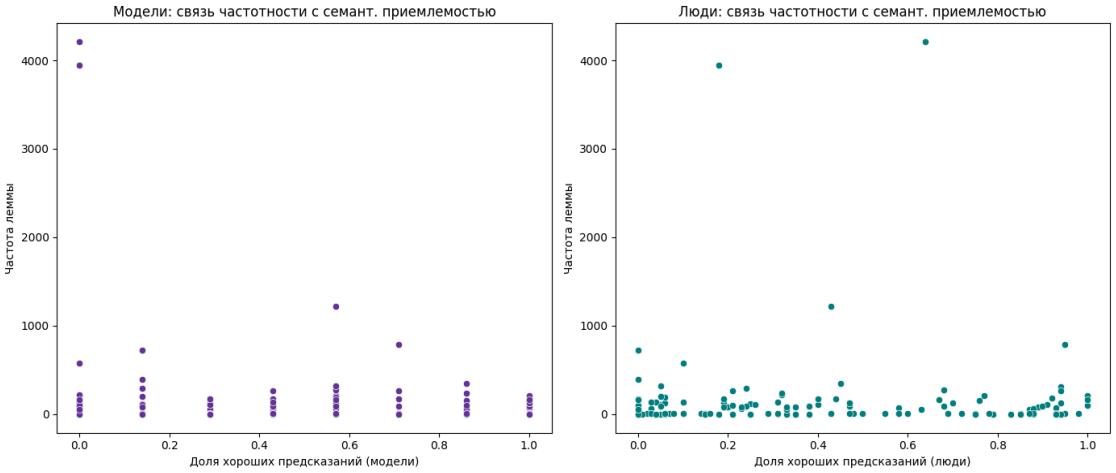
Данные результаты отчасти подтвердили **гипотезу 1**, показав что **длина таргетного слова влияет на точность предсказаний.**

8.5 Влияние частотности на предсказания

8.5.1 Влияние абсолютной частотности на приемлемость предсказаний

Как и для длины таргетного слова, была проверена гипотеза, что данный критерий может положительно влиять на семантическую приемлемость предсказаний. Критерий Пирсона для моделей равен -0.14, а p-value = 0.092, для людей эти значения равны 0.023 и 0.788, соответственно. То есть опять нет никакой линейной зависимости и различия статистически не важны. Этот вывод хорошо проиллюстрирован на рисунке (7). Видно, что все значения распределены, практически, равномерно по оси X.

Рисунок 7. Влияние абсолютной частотности целевого слова на семантическую приемлемость предсказаний.



Поэтому сначала мы рассмотрим влияние частотности леммы таргетного слова на точность предсказаний, а затем влияние частотности биграмм.

8.5.2 Влияние частотности (high / low) на точность предсказаний

Анализ точности предсказаний в зависимости от абсолютной частоты таргетного слова показал схожие результаты с выводами раздела 8.4.2. При сравнении точности и общей приемлемости предсказаний, люди оказались более чувствительны к частотности слов, чем модели (см. Рисунки 8 и 9), то есть есть зависимость, что чем слово частотнее, тем с большей вероятностью будет дано точное предсказание.

Рисунок 8. Зависимость точности предсказаний людей в зависимости от частотности таргетного слова.

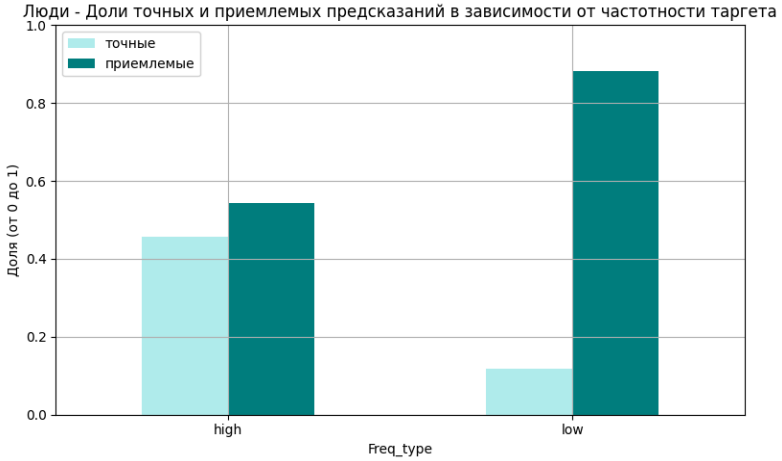
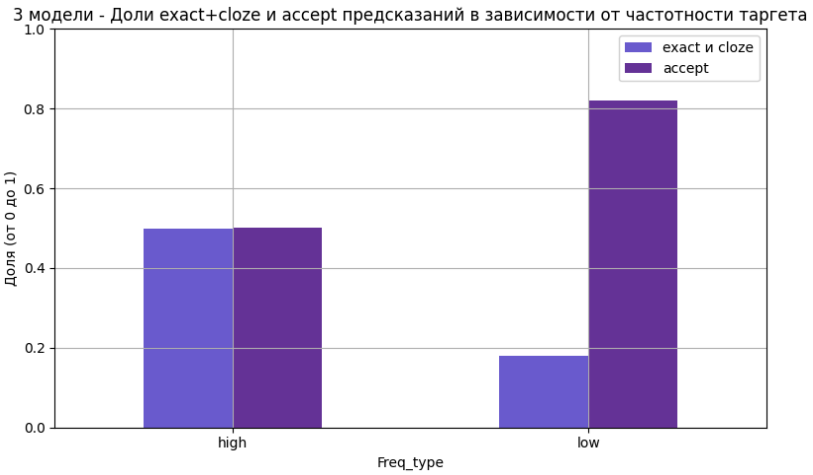


Рисунок 9. Зависимость точности предсказаний моделей в зависимости от частотности таргетного слова.



И среди предсказаний людей, и среди предсказаний моделей, доля точных предсказаний значительно выше для высокочастотных слов, чем для низкочастотных. У обоих типов испытуемых эта доля приближается к 0.5, то есть если целевое слово часто встречается в языке (вне зависимости от контекста предложения), то с 50% вероятностью, будет предсказано оно, а не какое-то другой слово, которое также может быть употреблено в рассматриваемом контексте.

При рассмотрении результатов моделей видно, что для высокочастотных слов доли предсказаний с метками Exact и Accept+Close практически равны, тогда как для низкочастотных слов разница достигает 60% – около 20% точных предсказаний и примерно 80% приемлемых.

Если говорить о людях, разрыв еще более заметен: около половины предсказаний - точные для высокочастотных слов, тогда как для низкочастотных точных предсказаний всего около 10%. Это подтверждает, что, подобно длине слова, частотность влияет на точность предсказаний сильнее у людей, чем у моделей.

8.5.3 Влияние частотности биграммы

Отметим, что при составлении исследуемых предложений никак не учитывалась частотность биграмм, которые образовываются целевым словом и предыдущим словом. Среди исследуемых предложений, есть такие, в которых предыдущее слово и таргет образовывают коллокацию, то есть это уже не просто набор двух слов, идущих друг за другом, а неслучайное сочетание двух лексических единиц, которые часто идут вместе (см. Пример (10)). При этом в большей части предложений никак очевидных коллокаций выделить нельзя (см. Пример (11)).

(10) Причиной аварии был [*мобильный <телефон>*], который отвлекал водителя от дороги.

(11) Применение [*микросхемы <дает>*] возможность уменьшить вес бытовой электроники.

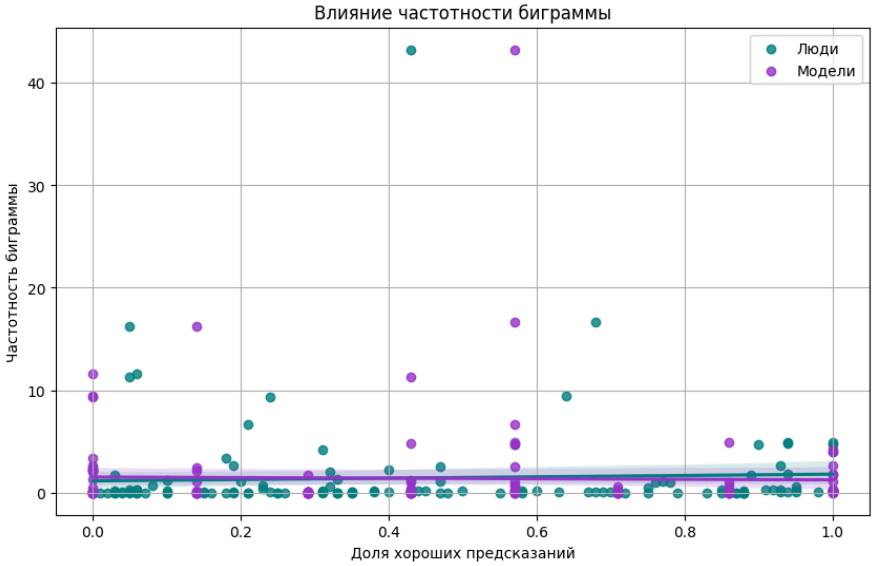
Для того, чтобы провести анализ, в каждом предложении были рассмотрены частотности биграмм, образованных целевым словом и предыдущим. В редком случае бралось не предыдущее слово, а чуть более отдаленное, как в примере (12). Затем в НКРЯ при помощи «лексико-грамматического поиска» искалось рассматриваемое сочетание, в качестве результата бралось значение параметра IPM[17]. Затем был построен график, представленный на рисунке (10). По нему видно, что большинство биграмм имеют значение IPM на уровне около 0, то есть в подавляющем числе предложений биграммы очень частотны. Отметим, что всего встретилось только 5 биграмм, частотность которых больше 10. Также по графику видно, что сильной корреляции между частотностью биграммы и долей приемлемых предсказаний нет. Это наблюдение подтверждают статистические метрики: p-value равно 0.8 и 0.55, а коэффициент Пирсона равен -0.02 и 0.05 для моделей и людей, соответственно.

В примере (12) было взято сочетание «раскрыть рот», потому что в данном предложении глагол «быть» является глаголом - связкой, и не выполняет самостоятельные функции. В целом, таких предложений, в которых были взяты не два подряд идущих слова, было очень мало, поэтому можно считать, что брались только биграммы в классическом определении этого понятия (два идущих подряд слова).

(12) Ваня [*раскрыл* (было) *<рот>*], но понял, что что-то не так, и промолчал..

[17] IPM ("Instances Per Million") в Национальном корпусе русского языка обозначает — количество вхождений данного слова или словосочетания на миллион слов корпуса

Рисунок 10. Влияние частотности биграммы на приемлемость предсказаний.



Возможно, такие результаты связаны с тем, что встретилось очень мало биграмм, значение IPM которых было бы больше 10.

8.5.4 Выводы из раздела 8.5

Как и с анализом длины таргетного слова, было установлено что **частотность не влияет на правильность предсказаний**, **нет линейной зависимости семантической приемлемости предсказаний от частотность биграмм**, образовываемых таргетным словом и предыдущем (см. раздел 8.5.3). В разделе 8.5.2 было показано, что есть небольшая **положительная зависимость между точностью предсказаний и частотностью целевого слова**. При этом **люди более чувствительны к увеличению** частотности слова, чем модели. Эти наблюдения также **частично подтверждают гипотезу 1**.

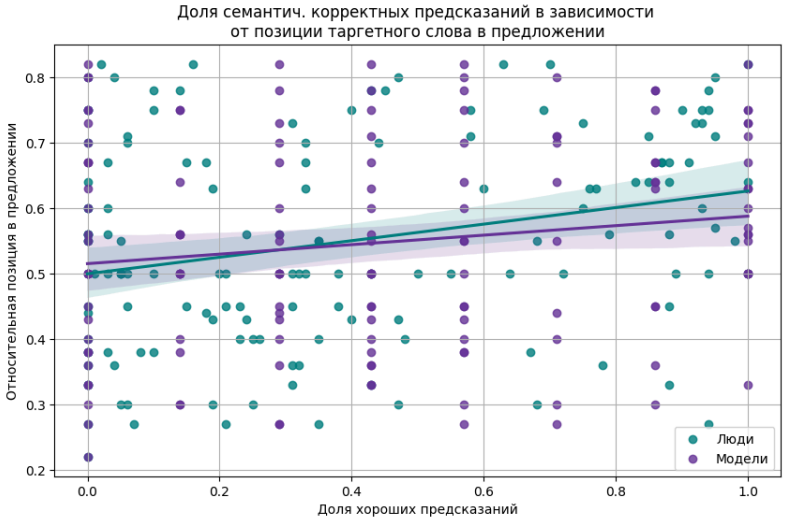
8.6 Влияние позиции таргетного предложения

С целью проверки гипотезы о том, что положение целевого слова в предложении коррелирует с приемлемостью предсказаний, для каждого предсказываемого слова была вычислена его удалённость от начала предложения, исходя из общей длины предположения. Предполагается, что с увеличением объёма видимого контекста будет возрастать вероятность приемлемого предсказания, как для моделей, так и для людей.

Рассмотрим рисунок 11. На нем показано, что между данными нет устойчивой сильной корреляции, но всё же присутствует небольшая зависимость: чем ближе слово

к концу предложения, тем с большей вероятностью предсказание будет семантически корректным. Чтобы подтвердить данное наблюдение был посчитан коэффициент Пирсона, чтобы проверить наличие линейной зависимости. Для предсказаний людей он равен 0.27, что указывает на слабую положительную взаимосвязь. Для предсказаний моделей коэффициент еще меньше и равен 0.16, что указывает на еще меньшую корреляцию. Можно сделать вывод, что и для моделей, и для людей нет четкой линейной зависимости между объемом видимого контекста и семантической приемлемостью предсказаний. Но этот параметр всё же немного влияет, при чем на людей в большей степени, чем на модели.

Рисунок 11. Зависимость приемлемости предсказаний моделей и людей в зависимости от относительной позиции таргетного слова в предложении.



8.6.1 Выводы из раздела 8.6

Результаты показали наличие небольшой **положительной корреляции между объемом видимого контекста и семантической приемлемостью предсказаний.** На людей этот фактор влияет в большей степени, чем на модели. Это отчасти подтверждает **гипотезу 2**.

8.7 Ошибки в управлении

Самая распространенная грамматическая ошибка (не считая синтаксических ошибок) - это ошибка в управлении. Для того, чтобы проанализировать их, сначала все предложения были отсортированы так, чтобы остались только те, в которых целевое слово – существительное. Затем были оставлены только те примеры, в которых в левом контексте содержится вершина таргетного слова (предлог, другое существительное или

глагол), что позволяет однозначно определить падеж, в котором должно стоять предсказание (см. примеры (13), (14) и (15)).

(13) Не поручайте *<мужу> (genitive)* ухаживать за рыбками в аквариуме, он обязательно забудет.

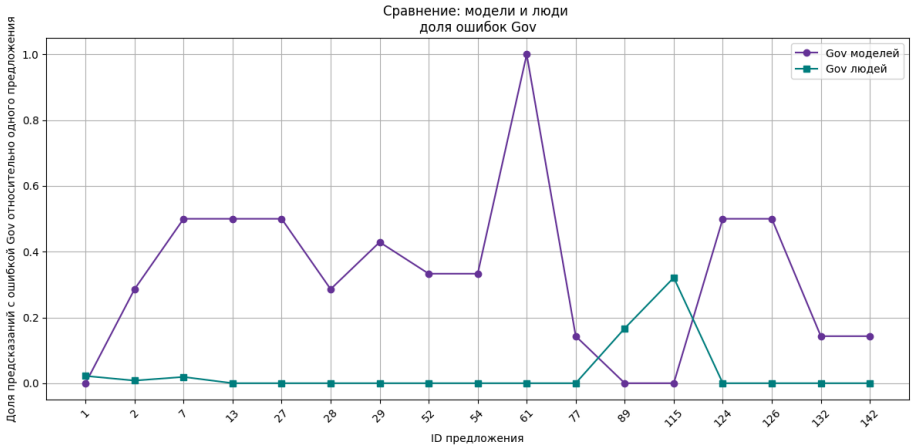
(14) Он умел из любого *<сора> (genitive)* извлекать информацию.

(15) Он ловко поддел концом *<ножа> (genitive)* замочки и они отскочили.

После сортировки для каждого предложения рассчитывалась доля предсказаний с ошибкой в управлении среди всех предсказаний, в которых в качестве ответа было выбрано существительное. Среди всех подходящих предложений люди ошиблись лишь в пяти предложениях из 69, модели в 14. При этом и люди, и модели ошиблись только в двух одинаковых предложениях (см. примеры (13) и (15)), остальные ошибки встретились в разных контекстах.

На рисунке (12) по оси X нанесены порядковые номера всех предложений, в которых были допущены ошибки, ось Y показывает долю предсказаний с ошибкой в управлении для каждого предложения. Заметим, что у людей нет ни одного предложения, в котором было бы предсказано больше половины слов, стоящих в неверной падежной форме. В то время как у моделей средняя доля ошибочных предсказаний гораздо выше. То есть этот график хорошо иллюстрирует гипотезу 4 о том, что если человек уже определился с той синтаксической стратегией, которой он будет следовать, то он допускает меньше грамматических ошибок, чем модель.

Рисунок 12. Доля предсказаний, стоящих в неправильной падежной форме



8.7.1 Выводы из раздела 8.7

Анализ ошибок в управлении показал, что **модели чаще ошибаются в выборе падежа**, подтверждая **гипотезу 4**.

8.8 Анализ лексических ошибок и ошибок в части речи

В данном разделе рассматривается гипотеза о том, что доля частеречных ошибок в предсказаниях может служить индикатором сложности предложения. Иными словами, предполагается, что чем сложнее контекст (в смысле предсказуемости), тем ниже доля приемлемых предсказаний, и тем выше вероятность ошибок в определении части речи. Это объясняется тем, что сделать успешное предсказание в пределах уже правильно угаданной части речи существенно легче, чем в рамках, когда часть речи еще не идентифицирована, так как испытуемому сначала необходимо «угадать» часть речи, а уже затем – лексему внутри предполагаемой группы. Именно поэтому ожидается, что лексические ошибки менее критичны, потому что круг возможных предсказаний уже сужен до конкретной части речи, остается только выбрать подходящую лексему.

Для отображения этой гипотезы было построено два графика по результатам работы моделей и людей (см. рисунок (13) и (14), соответственно). На обоих графиках изображены две линии тренда. Первая отображает зависимость доли приемлемых предсказаний от доли лексических ошибок. Вторая иллюстрирует изменение доли подходящих предсказаний в зависимости от доли частеречных ошибок. Важное уточнение: в этом анализе никак не учитываются грамматические ошибки (то есть «правильность» предсказаний рассматриваются только в рамках семантической разметки). Анализ наклона этих линий позволяет оценить, как различается чувствительность моделей и человека к рассматриваемым типам ошибок.

Рисунок 13. Lex и POS_Error, анализ предсказаний моделей.

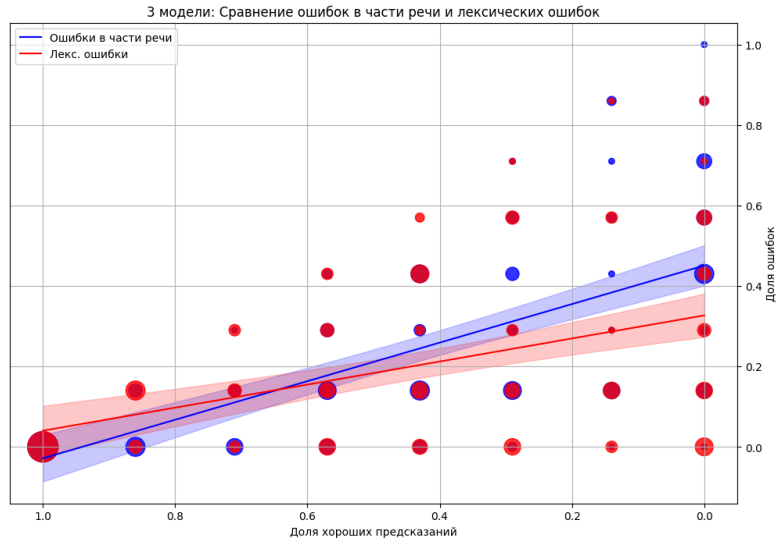
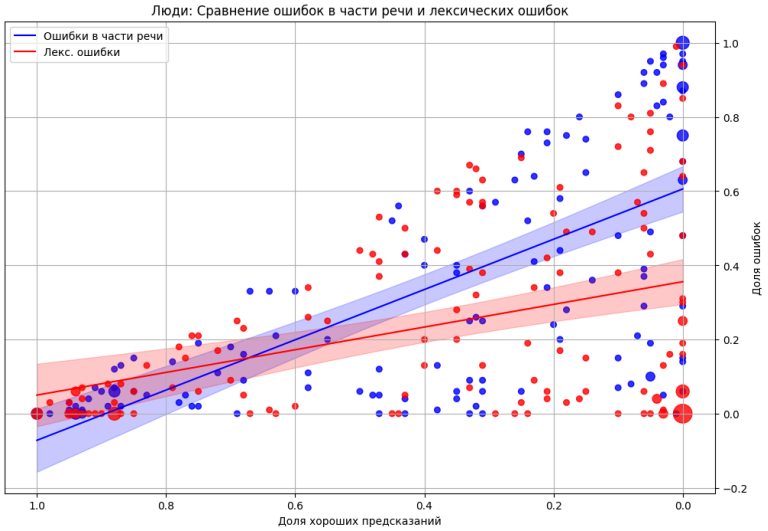


Рисунок 14. Lex и POS_Error, анализ предсказаний людей.



Проанализировав данные, представленных на двух графиках, можно сделать два ключевых вывода. Во-первых, люди и языковые модели демонстрируют схожую чувствительность к лексическим ошибкам: линии тренда для обоих графиков лежат в одном и том же диапазоне по оси Y — от 0.1 до 0.35. Во-вторых, поведение по отношению к частеречным ошибкам различается: линия тренда, отражающая зависимость приемлемости предсказаний от доли частеречных ошибок, у моделей имеет меньший наклон, чем у людей. То есть снижение количества семантически подходящих предсказаний при увеличении доли частеречных ошибок происходит у людей резче, чем у моделей. Это может еще раз подтверждать выводы, сформулированные ранее, о том, что люди более чувствительны к частеречным нарушениям.

8.8.1 Выводы из раздела 8.1

Анализ лексических и частеречных ошибок подтвердил **гипотезу 4**: частеречные ошибки действительно могут служить индикатором сложности предсказания. Было показано, что как у людей, так и у моделей **снижение доли приемлемых предсказаний сильнее коррелирует с увеличением частеречных ошибок, чем с лексическими**. При этом **у людей наблюдается более резкое снижение семантической приемлемости при росте доли частеречных ошибок**, что указывает на их повышенную чувствительность к ошибкам данного типа.

8.9 Отдельные замечания

8.9.1 Орфографические ошибки

Стоит заметить, что только среди предсказаний людей встретились слова с орфографическими ошибками, которые, видимо, связаны с опечатками, потому что эксперимент проводился онлайн и ответы записывались с клавиатуры, поэтому какие-то небольшие ошибки можно было не заметить или допустить случайно, например, телеффон, снех, хакки. Такие примеры помечались тегом Misspell. Важно отметить, что эти ошибки не должны были настолько исказить слово, что нельзя было бы восстановить, что имелось в виду (иначе должен был бы стоять тег Error). Всего таких примеров встретилось меньше 1% от всех предсказаний людей.

У людей встречались транслитерированные предсказания, например, uvidela, чего опять не было среди предсказаний моделей. Таких предсказаний было еще меньше, чем с орфографическими ошибками.

Также стоит отметить, что и у моделей, и у людей, были предсказания, которые нельзя никак интерпретировать, например, fff.

8.9.2 Анализ отдельных предложений

Помимо анализа по отдельным типам ошибок, также был проведен анализ по отдельным предложениям. Было выявлено, что есть предложения, в которых модели и люди ведут себя одинаково, например имеют высокий процент ошибок или наоборот дают много хороших предсказаний. При этом есть предложения, в которых поведение моделей и людей сильно отличается, одни предсказывают сильно лучше других. К сожалению, пока не удалось выяснить, что точно может влиять на различия в поведении. В этом разделе будут описаны только некоторые наблюдения, которые удалось получить на данный момент.

8.9.2.1 Предложения с наибольшим процентом приемлемых предсказаний

В 24 предложениях и модели, и люди получили очень высокие результаты, от 80% до 100% приемлемых предсказаний. Все эти контексты объединяет то, что целевые слова в них стоят во второй половине предложения, 80% таргетных слов – существительные. При этом по частотности скрытые слова распределены поровну (12 высокочастотных слов, 12 низкочастотных).

8.9.2.2 Предложения с наибольшим процентом ошибок в части речи

Можно выделить 10 предложений с наибольшим количеством ошибок в части речи, от 80% до 100% и у моделей, и у людей. Не совсем понятно, что может объединять эти примеры, но складывается ощущения, что просто сами конструкции, используемые в предложении сложнее для парсинга и менее предсказуемые. Рассмотрим несколько примеров предложений из этой группы. Рассмотрим предложения (16) и (17), мы скорее ожидаем после прилагательного увидеть существительное, а не еще одно прилагательное. В примере (18) мы видим, что нарушен канонический порядок слов SVO (subject, verb, object), сначала идет глагол, затем определение, затем существительное в роли агенса.

(16) Ей хотелось выплеснуть чай на бежевый <*Юлин*> пиджак, но она сдерживалась.

(17) Я люблю салат из картошки с зеленью, заправленный <*пахучим*> подсолнечным маслом.

(18) На газовой плите стояла <*открытая*> кастрюля с кипящей водой.

Большинство этих предложений объединяет характер предсказаний. Чаще всего были предсказаны слова, которые идут далее в предложении. То есть для (17) предложения часто предсказывались словоформы «маслом», «майонезом», «подсолнечным», «оливковым». Для (18) предложения наиболее частотным предсказанием было слово «кастрюля». То есть это всё слова, которые при наличии только левого контекста абсолютно грамматичны и приемлемы, но при анализе с правым контекстом становится понятно, что эти слова идут далее по контексту, и требовалось предсказать какой-то адъюнкт, а не аргумент.

8.9.2.3 Предложения с наибольшим процентом лексических ошибок

Среди всех предложений встретилось 4, в которых и модели, и люди, допустили наибольшее число лексических ошибок. Эти предложения объединяет то, что наиболее часто предсказываемые биграммы, состоящие из предсказанного слова и предыдущего, были в разы частотнее, чем изначальные биграммы (см. в таблице (5)). Но при этом предсказанные словосочетания никак не могут быть использованы в контексте целого предложения. В третьем столбце таблицы приведены биграммы, которые предсказывались сильно чаще остальных (то есть и люди, и модели часто предсказывали именно такие слова «нож», «нос» и т.д.).

Таблица 5. Сравнение частотностей биграмм.

Предложение	Частотность таргетной биграммы (IPM)	Частотность наиболее часто предсказываемой биграммы (IPM)
Ему удалось вскрыть банку об острый край <бампера> своего автомобиля.	край бампера - 0.01	край ножа - 0.04
У директора школы был тонкий <нюх> на талантливых педагогов.	тонкий нюх - 0.05	тонкий нос - 1.02
Под рукавом рубашки виднелся тонкий <ремешок> мужских часов.	тонкий ремешок - 0.15	тонкий шрам - 0.03
От смерти его спасла <собака>, приносившая ему еду.	спасла собака - 0.06	спасла случайность - 0.1

Возможно, нарушение этой тенденции в третьем предложении связано с тем, что тут на предсказуемость в большей степени влияет ассоциация с чем-то, что обычно скрывается, даже не обязательно под рубашкой, поэтому часто предсказывалось слово «шрам».

В этом же ключе интересно рассмотреть предложение (19). Среди предсказаний людей 63% – лексические ошибки, среди предсказаний моделей – 31% ошибок такого же рода. Наиболее частое неправильное предсказание - «кашу». Кажется, что промывать можно только крупу, а кашу можно варить. Не смотря на это, видимо, тут сыграл тот факт, что биграмма «манная каша» гораздо частотнее биграммы «манная крупа» (значение IPM 1.36 и 0.22, соответственно).

(19) А промывать манную <*крупу*> перед тем, как варить ее, не пробовали?

8.9.2.4 Предложения, в которых модели и люди ведут себя по-разному

Наконец, можно выделить 2 группы, в которых качество предсказаний моделей и люди сильно различается. Пока не получилось выделить какие-то факторы, которые могут объединять такие случаи. Далее будут продемонстрированы наиболее показательные предложения.

Рассмотрим примеры, с которыми люди справились сильно лучше чем модели ((20) и (21)):

(20) Выбирая вязаную шапочку, знайте, что лучше шапка цвета <*хаки*> из шерсти.

(21) Мне нравится сын коллеги, <*который*> недавно заходил в наш отдел.

У моделей в этих предложениях 100% ошибочных предсказаний (ответов с семантическими тегами Lex, POS_Error, Error), а у людей 64% и 75% предсказаний приемлемы, соответственно для каждого предложения.

А модели показали более высокие результаты в предложениях (22) и (23):

(22) Старуха была страшной -- <*лопоухой*> и с гнилыми зубами.

(23) Не поручайте <*мужу*> ухаживать за рыбками в аквариуме, он обязательно забудет.

У моделей тут больше 70% приемлемых предсказаний, а у людей всего лишь около 30%. Но непонятно, что может влиять на такой разброс в результатах. Этот вопрос планируется изучить при дальнейшей работе над этим исследованием.

8.9.5 Выводы из раздела 8.9

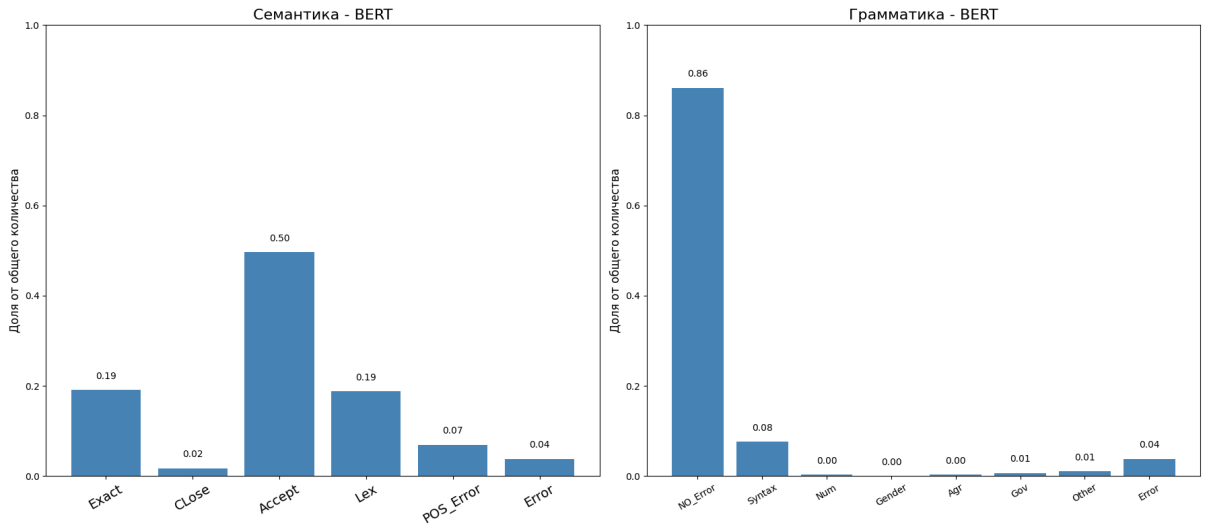
В разделе 8.9 было показано, что **есть предложения, в которых поведение моделей и людей различается**. Для понимания причин этого различия необходим дальнейший анализ структуры этих предложений. Также были показаны 2 типа ошибок, которые совершают только люди.

9. Анализ результатов экспериментов с BERT

До этого раздела ответы двух моделей BERT не были проанализированы, потому что в эксперименте были видны оба контекста, как левый, так и правый, что сильно облегчает саму задачу предсказания, делая ее уже даже не задачей предсказания, а задачей заполнения определенного пропуска.

В данном эксперименте были получены феноменально хорошие результаты (см. рисунок (15)). Больше чем 70% предсказаний приемлемы, при том, практически, 20% из них – точные попадания. Также 86% предсказаний грамматически верны. Было допущено пара грамматических ошибок, типа Num, Agr_ и Gov.

Рисунок 15. Общие результаты по семантической и грамматической разметкам.



Анализируя результаты, представленные в таблице (6), можно заметить, что тренд предсказуемости целевых частей речи сохраняется, и даже выражен в еще ярче. Предсказуемость прилагательных хуже, чем глаголов и существительных; грамматические характеристики предсказываются лучше, чем семантические (то есть предсказать семантически подходящие варианты тяжелее, чем грамматически правильные варианты). Разницу между верным попаданием в часть речи, при условии что целевое слово – существительное или глагол, можно считать статистически незначимой, потому что различие составляет меньше процента.

Таблица 6. Предсказуемость частей речи, BERT.

BERT			
	correct POS	correct sem	correct gram
noun	89,87	67,72	87,34
verb	90,91	74,24	86,36
adj	67,19	54,69	64,06

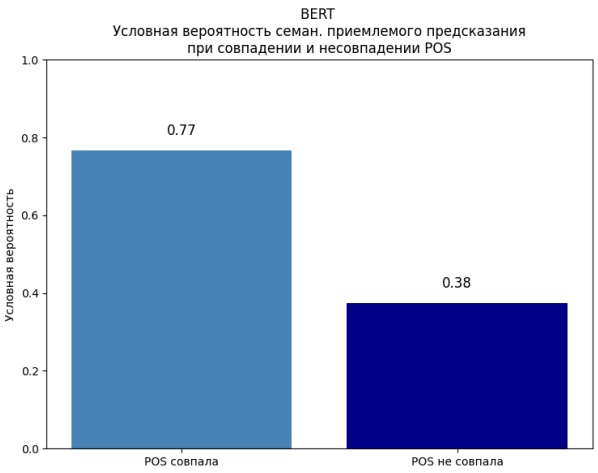
Рассмотрение приемлемости предсказаний, чья часть речи отличается от части речи целевого слова, показало, что прилагательные обладают максимальной заменяемостью (см. риснок (16)). Практически, в 60% случаев, когда вместо прилагательного предсказывалось слово другой части речи, это предсказание было приемлемым, как, например, в предложении (24):

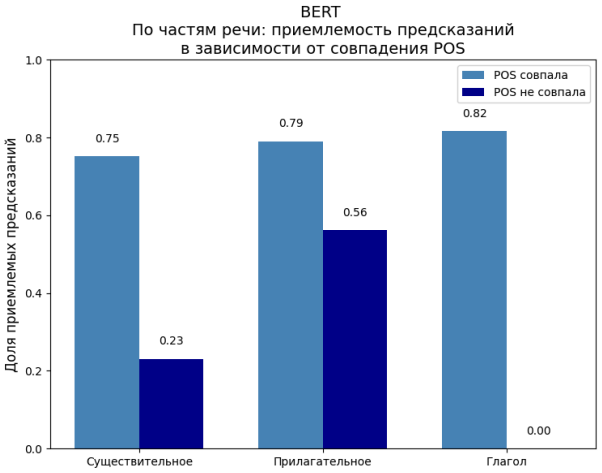
(24) Старый шкаф <*орехового*> дерева был явно не отсюда.

Предсказание: *из*

Также, процент случаев, когда вместо существительного было дано приемлемое предсказание другой части речи, выше, чем был у моделей и людей. Важно отметить, что при этом не было предсказано ни одного приемлемого слова другой части речи на месте глагола.

Рисунок 16. Соотношение приемлемых предсказаний в зависимости от попадания в ожидаемую часть речи, BERT.





9.1 Выводы из раздела 9

Как и предполагалось в **гипотезе 5**, благодаря двунаправленной архитектуре, **BERT показывает результаты сильно превосходящие результаты людей и моделей**: более 70% семантически приемлемых (из которых 20% с тегом Exact) и 86% грамматически верных предсказаний. Это подтверждает, что доступ к полному контексту значительно облегчает задачу и повышает точность предсказаний.

10. Результаты

Количественные результаты показали, что в рамках проведенного эксперимента люди демонстрируют более низкие показатели предсказуемости по сравнению с большими языковыми моделями (Qwen, Llama, GPT). Одним из возможных объяснений может быть различие в стратегиях предсказания: модели, как правило, выбирают наиболее частотные и простые варианты, тогда как люди склонны к более оригинальным и комплексным предсказаниям, которые часто предполагают конструкции, состоящие из нескольких слов (например, прилагательное + существительное или предлог + существительное). Однако поскольку эксперимент ограничен форматом одного слова, такие предсказания автоматически маркируются как некорректные из-за отсутствия продолжения, которое, в свою очередь, ведет к нарушению синтаксического строя изначального предложения. В связи с этим, помимо количественных результатов были также проанализированы качественные результаты, и было рассмотрено, как различные параметры могут влиять на предсказуемость, как моделей, так и людей.

В исследовании было установлено, что длина предсказываемого слова и его абсолютная частотность в языке оказывают влияние на точность предсказаний, но при этом почти никак не влияет на общую успешность предсказаний. Более короткие и более частотные слова предсказываются с большей точностью как людьми, так и моделями. Однако у людей этот эффект выражен сильнее.

Влияние частотности биграмм (сочетаний предсказанного слова и слова перед ним) оказалось минимальным. Несмотря на наличие коллокаций, их частотность, практически, никак не коррелирует с семантической приемлемостью предсказаний. Это, вероятно, связано с тем, что большинство исследуемых биграмм имели низкие значения IPM, а высокочастотных биграмм в выборке было недостаточно.

Относительное положение таргетного слова в предложении оказывает умеренное влияние на качество предсказания: ближе к концу предложения вероятность семантически приемлемого предсказания чуть выше. Однако линейная зависимость слаба и статистически значима лишь для предсказаний людей.

Люди совершают грамматические ошибки реже, чем модели. Были рассмотрены ошибки в управлении, средняя доля таких ошибок у моделей значительно выше, что говорит о недостаточной чувствительности моделей к синтаксической структуре предложения (рассматривались только те предложения, в которых вершина составляющей находится в левом контексте относительно целевого слова).

При анализе зависимости между двумя основными типами семантических ошибок и приемлемостью предсказаний выяснилось, что частеречные ошибки оказывают более сильное влияние на качество предсказаний, чем лексические. При этом у людей эта зависимость выражена ярче: снижение доли приемлемых предсказаний при увеличении количества частеречных ошибок происходит быстрее.

Ошибки в орфографии или, связанные с транслитерацией, встречались только в предсказаниях людей, но в очень малом количестве, поэтому они сильно не повлияли на общую картину.

Было обнаружено, что в ряде предложений поведение моделей и людей может совпадать, в ряде других – различаться. Пока не удалось найти однозначного набора факторов, который мог бы объяснить эти расхождения. Предполагается, что сложность

синтаксических конструкций и неканоничный порядок слов в предложении могут увеличивать количество ошибок, особенно у моделей. Отдельный интерес вызывают предложения, с которыми люди справляются сильно лучше, чем модели, и наоборот.

Эксперименты с моделями BERT, использующими как левый, так и правый контекст, продемонстрировали значительно более высокие показатели приемлемости и грамматической корректности. Это подтверждает, что наличие двухстороннего контекста значительно упрощает задачу предсказания. Тем не менее, даже в таких облегченных условиях наблюдается те же самые закономерности, что и для моделей с людьми: предсказания по прилагательным оказываются менее точными, чем по другим частям речи; предсказание грамматических характеристик легче, чем предсказание подходящей лексемы.

11. Библиография

1. Laurinavichyute, A.K., Sekerina, I.A., Alexeeva, S. *et al.* Russian Sentence Corpus: Benchmark measures of eye movements in reading in Russian. *Behav Res* 51, 2019. 1161–1178. URL: https://doi.org/10.3758/s13428-018-1051-6(дата обращения: 21.05.2025). Текст: электронный.
2. Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin, 124*(3), 1998. 372–422. URL: https://doi.org/10.1037/0033-2909.124.3.372(дата обращения 20.05.2025). Текст: электронный.
3. Ehrlich, S. F., & Rayner, K. Contextual effects on word perception and eye movements during reading. *Journal of Verbal Learning & Verbal Behavior, 20*(6), 1981. 641–655. URL: https://doi.org/10.1016/S0022-5371(81)90220-6(дата обращения 20.05.2025). Текст: электронный.
4. Richard Futrell, Ethan Wilcox, Takashi Morita, Peng Qian, Miguel Ballesteros, and Roger Levy. 2019. Neural language models as psycholinguistic subjects: Representations of syntactic state. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. pages 32–42. Minneapolis, Minnesota. Association for Computational Linguistics.
5. Jon Gauthier and Roger Levy. The neural dynamics of auditory word recognition and integration. 2023.

6. Taylor, W. L. Cloze procedure: A new tool for measuring readability. Journalism Quarterly, 30, 1953. 415-433.
7. Wilcox E. G. et al. On the predictive power of neural language models for human real-time comprehension behavior //arXiv preprint arXiv:2006.01912. – 2020.
8. Oh, B.-D., Schuler, W. Transformer-Based Language Model Surprisal Predicts Human Reading Times Best with About Two Billion Training Tokens // *Findings of the Association for Computational Linguistics: EMNLP 2023*. Singapore: Association for Computational Linguistics, 2023. C. 1915–1921. URL: https://aclanthology.org/2023.findings-emnlp.128/(дата обращения: 18.05.2025). Текст: электронный.
9. Merkx D., Frank S. L. Human sentence processing: Recurrence or attention? //arXiv preprint arXiv:2005.09471. – 2020.
10. Ashby J, Rayner K, Clifton C. Eye movements of highly skilled and average readers: differential effects of frequency and predictability. Q J Exp Psychol A. 2005 Aug;58(6):1065-86. doi: 10.1080/02724980443000476. PMID: 16194948.
11. Belinkov Y. et al. Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP //Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP. – 2024.
12. Lyu B. et al. Finding structure during incremental speech comprehension //ELife. – 2024. – T. 12. – C. RP89311.

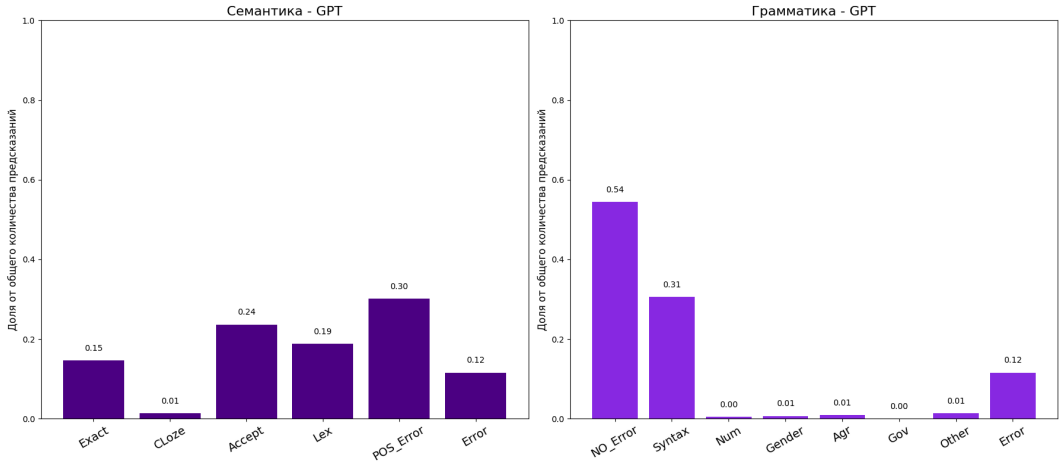
12. Приложение 1

```python
def predict_masked_word(sentence, tokenizer, model, max_length=3):
    if "<mask>" not in sentence:
        return ""

    # Разбиваем на левый и правый контекст
    prefix, _ = sentence.split("<mask>", 1)

    prompt = prefix.strip()
    inputs = tokenizer(prompt, return_tensors="pt").to(model.device)

    outputs = model.generate(
        **inputs,
        max_new_tokens=max_length,
        do_sample=True,
        top_k=20, # изменяемый параметр
        top_p=0.95 # изменяемый параметр
        temperature=0.7, # изменяемый параметр
        num_return_sequences=1,
        pad_token_id=tokenizer.eos_token_id,
        eos_token_id=tokenizer.eos_token_id
    )

    generated = tokenizer.decode(outputs[0], skip_special_tokens=True)

    predicted_part = generated[len(prompt):].strip()

    predicted_word = predicted_part.split()[0] if predicted_part else ""
    return predicted_word
```

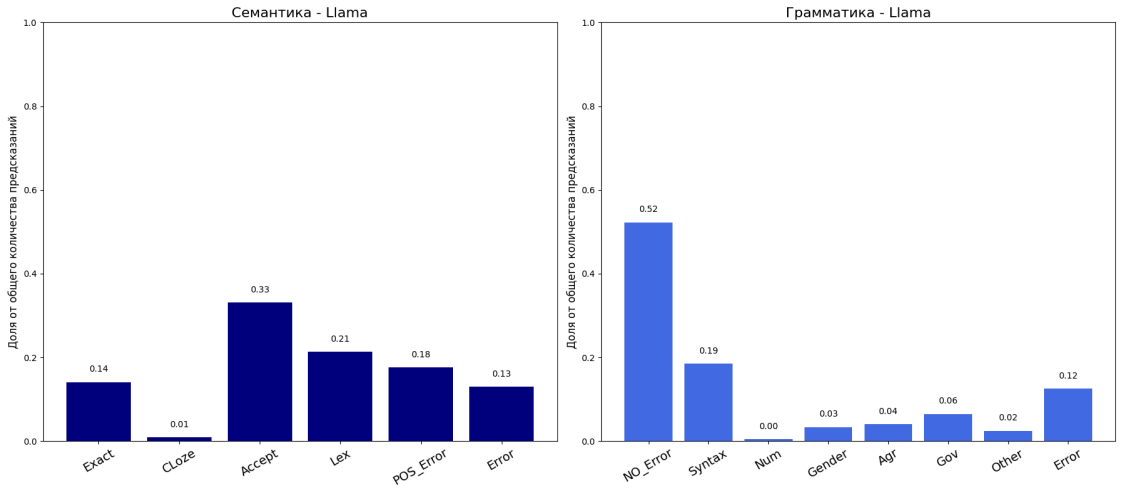
Полную версию кода можно найти по ссылке: https://github.com/VeraMonina/kursovaya_2025

13. Приложение 2

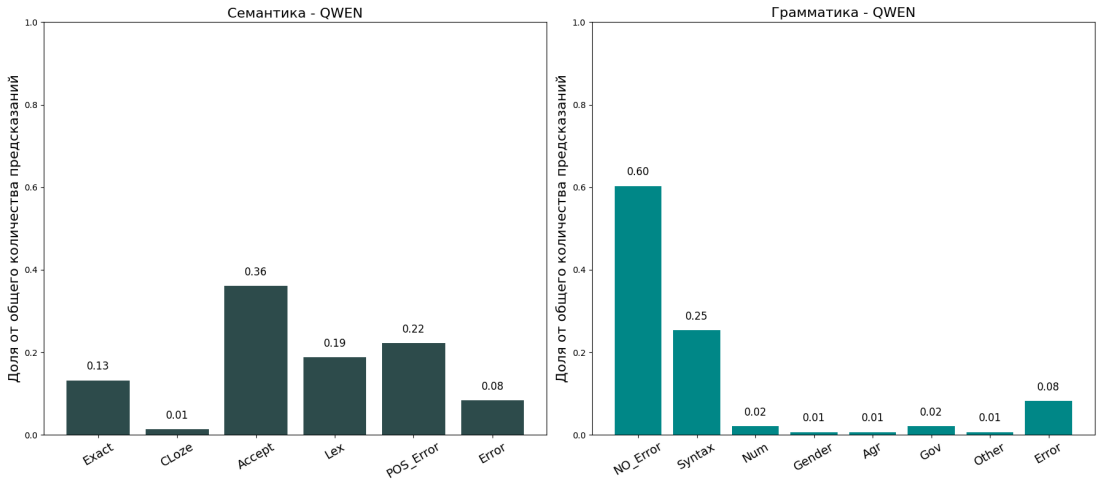
Результаты работы GPT:



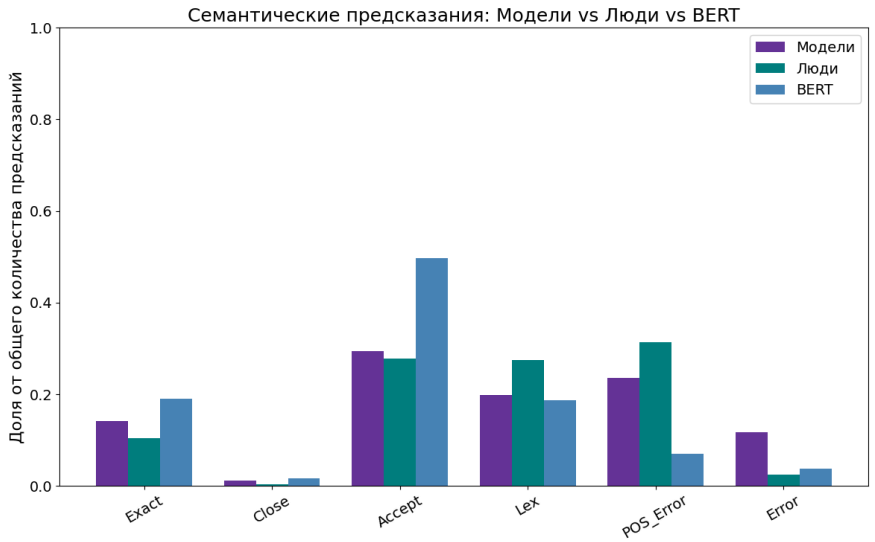
Результаты работы Llama:



Результаты работы Qwen:



14. Приложение 3



15. Приложение 4

Список 144 предложений, на которых производилось исследование, с выделенными в них таргетными слова.

1. В тот момент <*атмосфера*> наших посиделок внезапно сильно изменилась.
2. Клиенты воровали из ресторана <*атрибутику*> — скульптурки, кувшинчики, шкатулки.
3. Ему удалось вскрыть банку об острый край <*бампера*> своего автомобиля.
4. У моего отца был счёт в швейцарском <*банке*>, он был лесоторговец!
5. Очень хочется заплести <*бант*> в косу и надеть красивое платье.
6. В котел бросают куски <*баранины*>, специи, травы.
7. На запись голоса <*барса*> я наткнулся совершенно случайно.
8. Ее сын Гриша умер <*бездетным*>, и младший сын остался единственным наследником.
9. Можно будет <*бросить*> дополнительные занятия, и даже школу можно бросить.
10. Музыканты играли на похоронах, разгружали <*вагоны*>, жили бедно.
11. И не надо ставить это целью <*всей*> своей жизни.
12. Дрозды и скворцы начали <*вить*> семейные гнезда неподалеку друг от друга.
13. Что может сделать самый сильный <*вол*>, если он впряжен в сломанную повозку?
14. Здесь потребуется <*врач*>, который будет лечить заболевших в экспедиции?
15. Какие главные лекарства должны <*входить*> в аптечку автомобилиста?
16. Применение микросхемы <*даст*> возможность уменьшить вес бытовой электроники.
17. Выходя замуж, ты надеялась обрести спокойствие, уютный <*дом*>, крепкую семью.
18. С нескрываемой <*едкой*> иронией отзываются они друг о друге.
19. Однако здесь <*есть*> и свои актуальные проблемы.
20. Тому, кто <*желает*> невозможного, объяснят: Близок локоть, да не укусишь.
21. В качестве примера приводится <*жирность*> куриного бульона.
22. Он признаёт право каждого <*жить*> так, как ему удобно.
23. Вспоминая <*журчание*> водяных струй, мы непременно подумаем о фонтанах.
24. Он вскрыл пачку сухарей, <*заварил*> чай, достал чашки и ложки.

25. Но четыре года я не мог себя <*заставить*> сделать это.
26. На Ольгу Васильевну было написано <*заявление*> в полицию.
27. Я знал: их особенно <*злило*> то, что я никуда не бежал.
28. Очень тогда <*злые*> попадались люди.
29. Создать настоящие шедевры вам помогут <*зоркий*> глаз, терпение и упорство.
30. Наши власти позволяют себе <*излишества*>, создавая иллюзию благополучного общества.
31. У нас в Волгограде многие придерживаются <*иной*> точки зрения.
32. Ей никак не суметь <*испечь*> такой торт самой.
33. Во избежание ожогов надо нанести на лицо небольшое <*количество*> защитного крема.
34. В вопросе послышался упрёк <*командиру*>, словно он был виновником происшедшего.
35. За углом — Морской музей, с бесчисленными моделями <*кораблей*>, старинных и современных.
36. Мне нравится сын коллеги, <*который*> недавно заходил в наш отдел.
37. Душа требовала <*красоты*>, и Николай Фомин украсил свой дом резьбой.
38. Наше правительство сделало <*крен*> на дополнительные инвестиции в развитие науки.
39. А промывать манную <*крупу*> перед тем, как варить ее, не пробовали?
40. Каждое утро на самый верх <*крыши*> МГУ поднимается симпатичная научная сотрудница.
41. Убедительно просим вас разборчиво заполнять <*купон*>, желательно печатными буквами.
42. На болотах оставался ещё <*лёд*>, но на берегах реки появилась трава.
43. Дорога ведет в глухой <*лес*>, петляя по склонам.
44. На ведущей вниз <*лестнице*> сосед просил прекратить разговоры.
45. Когда она в самолёте <*летела*> домой, читать не было сил.
46. Возможности этих перемен будут обсуждаться в Париже <*летом*> будущего года.
47. Там, недалеко от кухонной двери, сидел <*лис*> с треугольной мордой.
48. В багажнике были лопата, <*лом*> и грабли.
49. Старуха была страшной -- <*лопоухая*> и с гнилыми зубами.
50. Врач прописал заживляющую <*мазь*> для обработки раны.
51. В бассейне <*микробы*> живут недолго, они привыкли к организму.
52. В резервациях <*миссионеры*> взялись их обращать в новую веру.
53. Мама брала меня с собой, и мы, сдав <*молоко*>, ехали гулять.
54. Приблизительно в центре тайги <*мопед*> упал в сугроб и сломался.
55. Не поручайту <*мужу*> ухаживать за рыбками в аквариуме, он обязательно забудет.
56. В деревнях по-прежнему <*мяли*> лен, дороги оставались непроезжими.
57. В числе возможных кандидатов <*называют*> депутата от партии правых.
58. Твоё тело расслабляется, и исчезает <*напряжение*> в области мышц.
59. Он был очень <*неопытным*> дипломатом и большим мечтателем.
60. Они не ели целый день, <*несчастные*> дети, и были очень голодны.
61. В речи учёного прозвучало <*неявное*> сопоставление с другим, знаменитым, физиком.
62. Тема эта в то время была <*новой*> для многих.
63. Существует легенда, что <*Ноев*> ковчег вынесло на вершину этой горы.
64. Он ловко поддел концом <*ножа*> замочки и они отскочили.

65. У директора школы был тонкий <*нюх*> на талантливых педагогов.
66. У Пашки <*обгорели*> ресницы и щеки, ходит, намазанный кремом от ожогов.
67. В конверте вместе с деньгами была <*обнаружена*> записка с угрозами.
68. Стала стабильнее экономическая и политическая <*обстановка*>, люди расслабились.
69. В современном <*обществе*> семья и школа оказывают большое влияние на подростков.
70. Педагог предъявляет <*одинаковые*> требования ко всем.
71. Я сказал, что русский солдат <*омоет*> сапоги в Индийском океане.
72. Старый шкаф <*орехового*> дерева был явно не отсюда.
73. В сюжете этого фильма какие-то <*осы*> устраивают себе гнёзда в дереве.
74. Ирине досталась <*отдельная*> комната в двухкомнатной квартире.
75. На газовой плите стояла <*открытая*> кастрюля с кипящей водой.
76. Я люблю салат из картошки с зеленью, заправленный <*пахучим*> подсолнечным маслом.
77. Ненужный коврик из твёрдой <*пластмассы*> пригодится как подставка для посуды.
78. Что ты хочешь чтобы тебе <*повторили*>, вопрос или ответ?
79. Журналист взял карандаш, <*подвинул*> к себе чистый лист и написал цифру.
80. У мамы есть <*подруга*>, которая живет прямо напротив здания театра.
81. Отвернув цветастое <*покрывало*>, она осторожно присела на край своей кровати.
82. У тебя впереди замечательный день, <*полный*> приятных событий.
83. Этот студент <*получает*> только отличные оценки с самого первого семестра.
84. Она почти не изменилась, только слегка <*пополнела*> и появились первые морщинки.
85. Перед ним снова была <*прежняя*> Маша, которую он знал и любил.
86. Торговля продуктами питания является одной из самых <*прибыльных*> отраслей в России.
87. Когда родители <*пригрозили*> не взять её с собой, Маша очень расстроилась.
88. Товарищ генерал, <*противник*> пробрался на наш командный пункт!
89. Считается, что коллекционирование <*раритетных*> машин — занятие для звезд шоу-бизнеса.
90. Телята быстро <*росли*>, превращаясь в ласковых коров.
91. Один футболист, который получил <*растяжение*>, не участвовал в игре.
92. Судя по огромному <*расходу*> воды, они купали слонов.
93. Город, раскинувшийся вдоль <*реки*>, состоял из двух частей.
94. Под рукавом рубашки виднелся тонкий <*ремешок*> мужских часов.
95. На вторичном рынке жилья <*розетки*> заклеивают обоями во время ремонта.
96. Ваня раскрыл было <*рот*>, но понял, что что-то не так, и промолчал.
97. В мои обязанности входило утром включить <*рубильник*> в подвале.
98. И на берегу озера тогда появляются <*русалки*> и ведьмы.
99. Наживка, на которую он ловил <*рыбу*>, быстро закончилась.
100. Не обнаружив ничего в досье, сыщики решили <*рыть*> в другом направлении.
101. Когда мне хотелось <*сгущенки*>, папа мне ее привозил, стоило только позвонить.
102. Взяв с собой фотоаппарат, вся <*семья*> поехала в парк на пикник.
103. Собаку, виновницу случившегося, приказали <*сечь*>, хотя в чем была ее вина?
104. Мне было лень идти на стоянку и сметать <*снег*> с машины.

105. От смерти его спасла <*собака*>, приносившая ему еду.
106. По воскресеньям музыканты, исполнявшие <*сонату*>, собирались в клубе.
107. Он умел из любого <*сора*> извлекать информацию.
108. Если я еще увижу здесь хоть <*соринку*>, я тебе уши оторву.
109. Количество денег в обороте выросло благодаря <*союзу*> с американским банком.
110. Он стал плохо <*спать*>, капризничать в детском саду.
111. Работы выполняет <*специалист*> от исполнителя, имеющий высокую квалификацию.
112. Зачем ему звонить, если откликается <*спокойный*> женский голос?
113. Я слезал, щупал <*стога*>, чтобы узнать, сухие ли.
114. Шею Лизы украшало ожерелье <*стоимостью*> в 15 тысяч долларов.
115. Этот роман захватывает читателя с первой <*страницы*> и держит до последней.
116. Автор принадлежит к числу последних свидетелей <*страшных*> событий прошлого века.
117. В темноте Иван задел острый <*сук*> и чуть не порвал рукав.
118. Приготовь себе диетические овощные блюда, <*сытные*> и восхитительные на вкус.
119. Олень бродил среди берёз, жевал <*талый*> снег и поглядывал за реку.
120. Причиной аварии был мобильный <*телефон*>, который отвлекал водителя от дороги.
121. После завершения <*техосмотра*> эксперты называют цену, по которой можно забрать автомобиль.
122. Чтобы придать объем <*тонким*> волосам, нанесите на них лечебную маску.
123. В лесу ветром <*трепало*> сухие стебли растений, почерневшие от летнего солнца.
124. В каждом <*углу*> комнаты сидели по две кошки.
125. Власть судов была такой <*узкой*>, что они ничего не решали.
126. Володя каким-то образом <*узнал*> то, чего ему не надо было знать.
127. За два года накопилась <*уйма*> вопросов, на которые надо было ответить.
128. Если мы позволим этим людям <*уйти*>, наши проекты станут гораздо беднее.
129. Думаю, большой <*урон*> здоровью такие вечеринки не наносят.
130. Елена сидела в кресле, молодая Мурка <*урчала*> у неё на коленях.
131. От внимания наблюдателя не должна <*ускользать*> даже малейшая деталь.
132. Государством предлагается <*установить*> ограничение на скорость движения.
133. Сделав мне знак помолчать, он приложил <*ухо*> к двери.
134. Она успевала убраться, разморозить <*фарш*> и приготовить ужин всем нам.
135. Что используют для этой прически, <*фен*> или ещё что?
136. Выбирая вязаную шапочку, знайте, что лучше шапка цвета <*хаки*> из шерсти.
137. Она с досадой <*хмурила*> брови, оглядываясь на дом.
138. Зоопарк — это кусочек другого мира, находящийся в самом <*центре*> нашего района.
139. Я сделала <*шаг*> навстречу: приехала к ней, попросив выслушать меня.
140. За министром труда тянется целый <*шлейф*> финансовых скандалов.
141. Мы установили камеру на новый <*штатив*> и приступили к съемкам.
142. Ей хотелось выплеснуть чай на бежевый <*Юлин*> пиджак, но она сдерживалась.
143. На привале у озера <*юный*> турист вручил товарищам ведро и кувшин.

144.Покуда я нахожусь у власти, я буду предметом <*ярых*> нападок соперников.