

УДК 81-13

DOI 10.17726/phillT.2023.1.4



Стратегии перевода машинных погрешностей в автоматически сгенерированных текстах (на примере GPT-4)

Алейникова Валерия Игоревна,

студентка 2-го курса магистратуры

Института переводоведения, русистики и многоязычия,

Пятигорский государственный университет,

Пятигорск, Россия

Aleinikova1405@gmail.com

Аннотация. В статье рассматриваются стратегии перевода «машинных текстов» на примере генеративных трансформеров (GPT). В настоящее время изучение и разработка генерации машинных текстов стали важной задачей для обработки и анализа текстов на разных языках. Современные технологии искусственного интеллекта и нейронных сетей позволяют создавать мощные инструменты для деятельности в этой сфере, которые с каждым годом становятся все более эффективными. Одним из таких инструментов являются генеративные трансформеры. Изучение генеративных трансформеров также позволяет разработчикам создавать более точные и эффективные алгоритмы машинного перевода, что повышает качество переводов и улучшает пользовательский опыт. В этом контексте особого интереса заслуживают особенности машинных текстов, созданных генеративными трансформерами, их закономерности, погрешности и несовершенства, которые требуют особых стратегий перевода. Сегодня можно говорить о том, что генерация уникальных и релевантных текстов – это рутинная задача, которую удалось автоматизировать. Тем не менее, определенные ограничения для использования таких текстов все еще существуют, в частности их использование требует применения соответствующих стратегий перевода. В работе предложена авторская типология стратегий перевода, куда с учетом особенностей АСТ предложено добавить подстратегию терциарно-модерационного перевода.

Ключевые слова: стратегии перевода; машинный перевод; автоматически сгенерированные тексты; машинные погрешности перевода; генеративные трансформеры.

Strategies for translating machine errors in automatically generated texts (using GPT-4 as an example)

Aleynikova Valeria Igorevna,

2nd year graduate student

*of the Institute of Translation Studies, Russian Studies and
Multilingualism of Pyatigorsk State University,
Pyatigorsk, Russia*

Aleinikova1405@gmail.com

Abstract. The article discusses the strategies of translation of «machine texts» on the example of generative transformers (GPT). Currently, the study and development of machine text generation has become an important task for processing and analyzing texts in different languages. Modern technologies of artificial intelligence and neural networks allow us to create powerful tools for activities in this field, which are becoming more and more effective every year. Generative transformers are one of such tools. The study of generative transformers also allows developers to create more accurate and efficient machine translation algorithms, which improves the quality of translations and improves the user experience. In this context, the features of machine texts created by generative transformers, their patterns, errors and imperfections, which require special translation strategies, deserve special interest. Today we can say that the generation of unique and relevant texts is a routine task that has been automated. Nevertheless, certain restrictions for the use of such texts still exist, in particular, their use requires the use of appropriate translation strategies. The paper proposes the author's typology of translation strategies, where, taking into account the features of AST, it is proposed to add a substrategy of tertiary-moderation translation.

Keywords: translation strategies; machine translation; automatically generated texts; machine translation errors; generative transformers.

Процесс перевода предполагает выстраивание определенной последовательности действий – стратегии, согласно которой будет осуществляться перевод. Долгое время понятие перевода сводилось к созданию полного эквивалента текста ИЯ на ПЯ, особое внимание уделялось при этом именно лингвистическим факторам. Тем не менее любой текст несет в себе определенную цель, которую необходимо достичь при акте коммуникации.

В исследованиях В. М. Илюхина стратегия перевода представляется как «метод выполнения переводческой задачи, заключающийся в адекватной передаче с ИЯ на ПЯ коммуникативной интенции отправителя с учетом культурологических и личностных особенностей оратора, базового уровня, языковой надкатегории и подкатегории» [1, с. 206].

Важными факторами, определяющими специфику коммуникативных ситуаций, считаются инициатор перевода и цель перевода. Многообразие переводческих стратегий побуждает переводоведов к попытке их систематизации. Изучая алгоритм переводческой деятельности, В. Н. Комиссаров пришел к выводу, что важным критерием выбора переводческой стратегии является знание параметров коммуникативной ситуации. В. В. Сдобников утверждает, что было бы более корректно подразделять стратегии на типы не согласно видам перевода (устный / письменный), а в зависимости от разнообразия коммуникативных ситуаций. Автор аргументирует свою точку зрения тем, что использование одного и того же вида перевода в разных коммуникативных ситуациях не означает, что в этих КСП используются одни и те же стратегии перевода, которые будут, несомненно, разные, так как это зависит от характера самой коммуникативной ситуации, в которой используется перевод [2, с. 464].

В ситуации, когда мы имеем дело с автоматически сгенерированным текстом, обеспечение адекватности перевода, подразумевающей условие установления взаимопонимания между разноязычными коммуникантами, то есть «обеспечения коммуникативной, интерактивной и перцептивной сторон общения» [3, с. 284], теряет смысл, поскольку такая адекватность недостижима за неимением автора оригинала, обладающего определенными коммуникативными намерениями. Соответственно, о стратегии коммуникативно-равноценного перевода здесь речи не идет. Мы также не в полной мере можем говорить о стратегии переадресации, поскольку изначальной адресации при создании текста машиной тоже не было. Алгоритм перевода АСТ ближе всего к стратегии терциарного перевода.

Однако следует учитывать фактор, который не всегда присутствует при переводах, имеющих стандартную схему человеческой коммуникации между автором текста и реципиентами оригинала и перевода: при переводе АСТ должна учитываться необходимость обязательной модерации текста переводчиком по таким критериям современного автоматически сгенерированного текста, как наличие

фактических ошибок, правильность грамматики и орфографии, верно выбранные языковые средства, включая стилистику и тональность, точность распознавания контекста темы (включая социальный, этнический, религиозный, географический и т.д. контекст), этичность (как сейчас пишут, «уровень токсичности») текста, его безопасность в плане использования конфиденциальных данных. В связи с этим стоит выделить в стратегии терциарного перевода подвид стратегии, связанный именно с переводами АСТ: стратегию терциарно-модерационного перевода, учитывающего проверку и редактирование АСТ по всем указанным критериям (рис. 1). Такая стратегия должна применяться и при создании перевода АСТ переводчиком-человеком, и при подготовке АСТ к машинному переводу.



Рисунок 1. Предлагаемая типология стратегий перевода с учетом особенностей АСТ

Схематично подстратегию терциарно-модерационного перевода можно изобразить следующим образом (рис. 2).

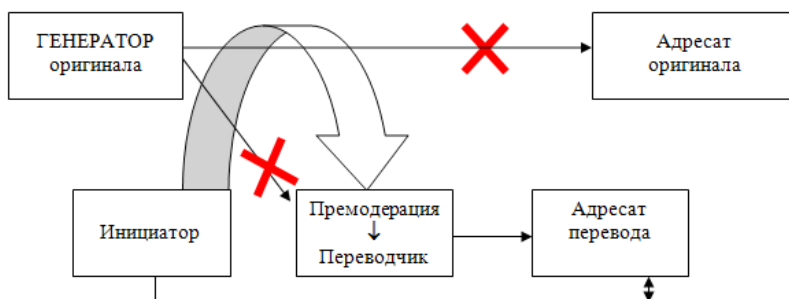


Рисунок 2. Подстратегия терциарно-модерационного перевода

Изучение генерируемых моделью ChatGPT текстов позволило выявить разнообразные погрешности, требующие внимания при переводе. Так, можно отметить лексические, стилистические погрешности, использование анахронизмов, например синонимов, которые не совсем соответствуют описываемой эпохе в тексте исторической направленности:

*Монголо-татарское нашествие на Русь началось в 1223 году, когда монгольские войска победили русских князей на реке Калке. После этого монголы начали захватывать и покорять все большие и больше территорий Руси, организовывая регулярные **налеты** и требуя покорности от русских князей.*

Слово «налеты» в данном контексте можно считать лексической погрешностью, поскольку оно не соответствует рассматриваемой эпохе монголо-татарского нашествия на Русь. Согласно Национальному корпусу русского языка, слово «налет» в значении «нападение» используется только с 1812 г. (а до этого – в 1739 г. только в контексте прилета птиц) [4] (рис. 3). С татаро-монгольскими нападениями чаще ассоциируется слово «набег».

Можно встретить и примеры грамматических ошибок, например несогласованности форм слов в русском тексте:

*Нельзя сказать, что **русские народ** угрюмее или менее веселый, чем другие народы. Каждый народ имеет свой уникальный характер и культуру, которые могут отличаться в зависимости от многих факторов, таких как климат, история и социальные условия.*

*Однако, есть стереотип, что **русские народ** угрюмы и склонны к меланхолии. Этот стереотип может быть связан с тем,*

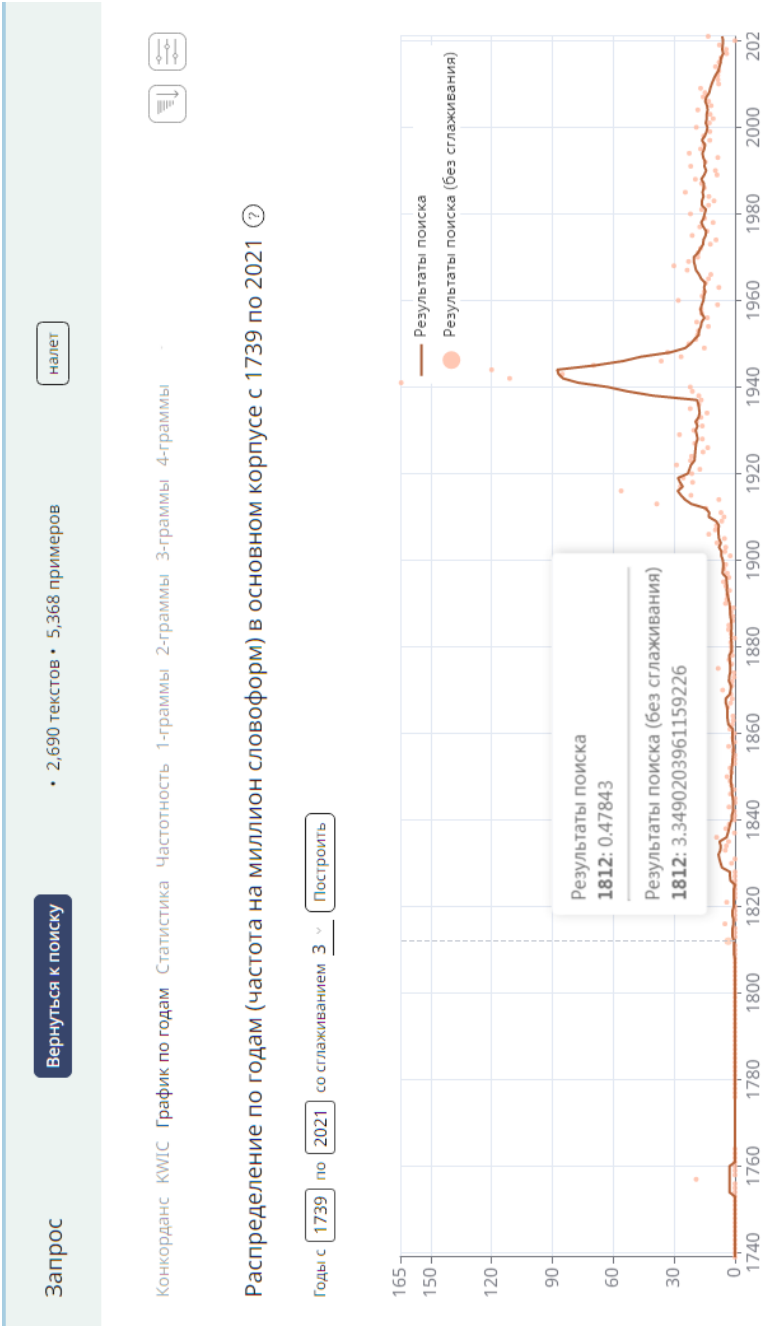


Рисунок 3. Слово «Налет». Распределение по годам

что русская культура и литература славятся своими глубокими и трагическими произведениями, которые выражают многие темы, связанные с жизнью и смертью, любовью и страданиями.

Как видно из примера, сочетание «русские народ» не разовая ошибка, а устойчивый результат самообучения машины на примере какого-то ошибочного текста, попавшего в базу машинного обучения. При переводе, особенно если предполагается дальнейший машинный перевод, это может стать причиной неверно переведенной фразы.

Грамматические и лексические ошибки можно увидеть и в следующем примере текста, автоматически сгенерированного GPT:

*Также, стоит отметить, что в итальянской культуре существует традиция многократного обнимания и **поцелуев на щеках** при приветствии и прощании, что может быть необычным для людей из других культур. Однако, это просто проявление итальянской традиции и обычая.*

*Однако, как и в любой культуре, **эмоциональность** итальянцев **могут** иметь свои недостатки и проблемы. Например, иногда это может приводить к более ярким конфликтам и спорам, чем в других культурах. Также, это может привести к тому, что итальянцы часто проявляют эмоции без фильтрации, что может быть неприемлемым в других культурах.*

*Таким образом, можно сказать, что итальянцы действительно **являются** очень эмоциональным **народом, которые** проявляют свою страсть и чувства во многих аспектах жизни. Это делает их культуру уникальной и привлекательной для многих людей со всего мира.*

Выражение «поцелуй на щеках» в данном контексте ошибочно, в русском языке используется «поцелуй в щеку», ни в коем случае не «на щеку». Заметны и погрешности в согласованности числа («эмоциональность» – ед. ч., «могут» – мн. ч.). Фраза «являются народом, которые...» тоже отмечена той же ошибкой, очевидно, что машина не всегда распознает разницу между словами «народ» и «люди».

В качестве не совсем этичного, «токсичного» текста можно привести такой отрывок:

*Однако, как и в любой культуре, **педантичность немцев** может иметь свои недостатки и проблемы. Например, она может приводить к тому, что немцы **часто уделяют большие внимания деталям, чем главным вещам, и это может приводить к задержкам или проблемам в работе.** Также, **чрезмерная педантичность может привести к тому, что немцы не смогут наслаждаться***

жизнью и расслабляться, что может повлиять на их здоровье и благополучие.

Любой представитель немецкого народа может счесть этот отрывок по меньшей мере задевающим или обидным. И хотя машина обучена сглаживать категоричность подобных суждений оборотами «может приводить» вместо «приводит», тем не менее токсичность подобного текста этим полностью не снимется.

В качестве примера фактической ошибки приведем текст юридической направленности:

Налог на наследование в России действительно существует, однако, размер налога зависит от многих факторов, в том числе от типа наследства и степени родства наследников с наследодателем.

*Стоит отметить, что в России нет специального налога на недвижимость, полученную в наследство. Налог на наследство взимается только в случае, если сумма наследства превышает пороговое значение, которое устанавливается законом. **С 2021 года в России налог на наследство начинается с суммы наследства свыше 5 миллионов рублей, при этом наследники первой очереди (супруг, дети) освобождаются от уплаты налога на долю в наследстве до 15 миллионов рублей.***

Таким образом, налог на недвижимость, полученную в наследство, в России не является чрезмерно высоким, и его размер зависит от многих факторов, в том числе от степени родства наследника с наследодателем и суммы наследства. Кроме того, налог может быть существенно снижен или даже полностью освобожден, если наследник будет использовать недвижимость в личных целях.

Учитывая, что как таковой налог на наследство в России, как известно, не существует (налог с продажи наследства – это уже несколько иная форма налога), а приведенные цифры и даты не имеют ничего общего с какими бы то ни было российскими законодательными новациями в сфере налогообложения наследников, данный АСТ является фактически ошибочным, что следует учитывать тем, кто по какой-либо причине намерен сделать его перевод. В целом очевидно, что юридическая тематика может быть наиболее уязвимой сферой использования автоматических систем генерации текстов, даже таких передовых, как GPT. Возможно, в такой ситуации потребуются комментарии переводчика с пояснениями относительно реального положения вещей.

В следующем примере можно увидеть, как машина может спутать в русском языке члены предложения, которые могут быть выражены одним словом, но в рамках одной фразы сочетаться не могут:

*Как искусственный интеллект, я не могу иметь друзей в обычном смысле, потому что я не обладаю личностью и не могу чувствовать эмоции. **Однако**, я предназначен для помощи людям, и я всегда готов помочь с решением проблем и ответами на вопросы, если это в моих возможностях. Моя задача – быть **полезным инструментом для людей, а не друзьям**.*

Здесь очевидно спутаны функции определения («**полезный** инструмент») и составного сказуемого («быть **полезным** кому-либо»), поэтому фраза вышла неверно составленной, части предложения между собой не согласованы. Здесь можно отметить и другую ошибку, как и во многих приведенных примерах, – ошибку синтаксическую: «однако» в значении «но» в начале предложения в русском языке не обособляется, но в текстах GPT это ошибочное обособление встречается постоянно.

Погрешности в понимании контекста можно отметить и на примере использования мемов. Так, в вопросе «Вы продаете рыбов?» (он стал одним из самых известных мемов 2021 года [5]) машина распознать мем не смогла и дала ответ без учета актуального сетевого культурного контекста:

*К сожалению, я не могу продать рыбу или какой-либо другой товар, так как я являюсь компьютерной программой, созданной для обработки текстовой информации и предоставления ответов на вопросы. Моя задача – помочь вам найти ответы на ваши вопросы и решить ваши проблемы, используя мои знания и возможности работы с текстом. **Если у вас есть вопросы о рыбе, я с радостью помогу вам найти информацию об этой теме.***

Генерация текстов на английском языке дает меньше ошибок, поскольку рассматриваемые модели-трансформеры изначально обучались именно на английских текстах. Тем не менее, и в английских текстах порой можно встретить те или иные погрешности. Например, грамматические:

*The thing **he was worried the most** was his finances. Money was a constant source of stress for him, and he often found himself struggling to make ends meet. He worked long hours to try and make more money, but it never seemed to be enough. He worried about paying his bills on time, saving for the future, and being able to provide for his family.*

Despite his best efforts, he always felt like he was one financial setback away from disaster. His worries about money consumed his thoughts and made it difficult for him to enjoy life. Более правильно было бы написать «The thing **he was most worried about** was his finances».

Можно в определенных ситуациях столкнуться и с лексическими недочетами. Например, в тексте «*she stopped eating **junks**. She realized that the unhealthy foods she was consuming were making her feel sluggish and unwell, and she decided to make a change. She started by cutting out sugary snacks, fast food, and processed foods, and instead focused on eating whole, nutritious foods like fruits, vegetables, and lean proteins*» слово **junks** используется не совсем правильно. Здесь речь, судя по контексту, идет именно о junk food, то есть о нездоровой, вредной еде, а **junks** означает чаще всего мусор, иногда наркотики и т.п. Если переводить эту фразу через программы машинного перевода, например через переводчик Google, то будет получен текст «она перестала есть всякую всячину», что явно не отражает общий контекст о нездоровой еде и фастфуде, поскольку «всякая всячина» в русском языке не всегда несет негативную коннотацию и не равна понятию «вредная еда».

Если говорить о премодерации АСТ на английском языке, особенно в том случае, если предстоит машинный перевод, то можно говорить о следующих ограничениях такого рода переводов. В первую очередь стоит упомянуть, что поощряется использование функциональных слов, таких как артикли, в то время как не приветствуется использование местоимений и союзов, т.к. они повышают синтаксическую многозначность. Также необходимо сократить использование причастных форм (на -ing и -ed) после союзов (*While driving the vehicle...) или в сокращенных определительных придаточных предложениях (*Directional stability caused by wheel lock-up...). Оба предложения должны быть переписаны следующим образом: While you are driving the vehicle... и The directional stability that is caused by the wheel lock-up.

Использование грамматических ограничений рекомендуется даже в тех случаях, когда текст не предназначен для перевода. Это связано с тем, что такие ограничения повышают удобочитаемость текста и снижают его многозначность. Ограничения бывают двух типов: связанные с отдельными фразами или с конструкцией всего предложения в целом. К фразовым ограничениям относятся, например, замена фразовых глаголов однословными (например, замена

«turn on» на «start») и избегание повторяющихся предлогов в соединительных конструкциях для предотвращения двусмысленности.

Например, в следующем тексте: *«Closed-circuit television (CCTV) systems are widely used in public places, workplaces, and homes to enhance safety and security. These systems use cameras to capture footage in specific areas, which is recorded and stored for later use.*

The recorded memory of surveillance system and every CCTV unit can provide valuable evidence in the event of accidents, incidents, or crimes, and can help identify suspects and gather information to solve cases.

CCTV systems are becoming increasingly sophisticated, with features like facial recognition technology, motion sensors, and remote access becoming more common. However, the use of CCTV systems also raises concerns about privacy and civil liberties, as constant monitoring of individuals and public spaces can lead to an erosion of privacy rights» – фраза «recorded memory of radio and each control unit» может рассматриваться как имеющая два значения:

- {recorded memory of surveillance system} and {every CCTV unit},
- recorded memory of {surveillance system and every CCTV unit}.

Ее стоит записать иначе: «recorded memory of the surveillance system and of each CCTV unit».

Рассматривая ограничения, относящиеся ко всей конструкции предложения в целом, стоит учитывать нюансы, связанные с сочинительным соединением частей предложения (в сложносочиненном предложении обе части должны относиться к одному типу), при подчинительном соединении – особенности использования придаточных предложений с функцией определения (их необходимо ввести в предложение относительным местоимением), а также важно обращать внимание на предложения, использующие эллипсис, – такие конструкции лучше вообще исключать из текста и заменять полными предложениями.

Выводы

Изучая понятия стратегии перевода с позиций лингвистического и коммуникативно-функционального подходов, можно выделить два основных направления в определении данного термина. С лингвистической точки зрения стратегии перевода рассматриваются как способы решения переводческих проблем. В этом подходе акцент делается на языковых аспектах, таких как эквивалентность, грамматика, лексика и другие языковые единицы. Переводчик выбирает

стратегию, которая наиболее точно передаст смысл и структуру исходного текста на целевой язык. С коммуникативно-функциональной точки зрения стратегии перевода направлены на передачу функции текста и его цели, учитывая параметры коммуникативной ситуации. Здесь важно понимание цели исходного текста, намерения автора, аудитории, для которой предназначен перевод. Переводчик выбирает стратегию, которая наилучшим образом соответствует коммуникативной задаче и потребностям целевой аудитории.

При изучении стратегий перевода в рамках коммуникативно-функциональной теории следует обратить внимание на то, что каждая стратегия представляет собой модель поведения переводчика и включает в себя определенные элементы. Эти элементы могут варьироваться в зависимости от конкретной стратегии и коммуникативной ситуации. Инициатор перевода и цель перевода являются важными факторами, определяющими специфику коммуникативной ситуации. Инициатор перевода может быть самым отправителем исходного текста, переводчиком или получателем целевого текста. Каждый из них имеет свои цели и ожидания от перевода, что может повлиять на выбор стратегии перевода.

В процессе работы над текстом переводчик проводит анализ исходного текста, выявляет его особенности, а затем принимает решение о выборе конкретной стратегии перевода. Этот анализ помогает переводчику понять цели и коммуникативные намерения автора исходного текста, а также адекватно передать их в целевом тексте. В целом выбор стратегии перевода осуществляется на основе анализа исходного текста, выявления его особенностей и учета целей инициатора перевода. Это позволяет переводчику наилучшим образом передать коммуникативную интенцию и особенности исходного текста в целевом тексте.

Существуют разные типологии переводческих стратегий. Рассмотрена типология стратегий перевода, включающая: 1) стратегию коммуникативно-равноценного перевода; 2) стратегию терциарного перевода; 3) стратегию переадресации. Сделан вывод, что в ситуации, когда мы имеем дело с автоматически сгенерированным текстом, обеспечение адекватности перевода, подразумевающей условие установления взаимопонимания между разноязычными коммуникантами, то есть обеспечения коммуникативной, интерактивной и перцептивной сторон общения, теряет смысл, поскольку такая адекватность недостижима за неимением автора оригинала,

обладающего определенными коммуникативными намерениями. Соответственно, о стратегии коммуникативно-равноценного перевода здесь речи не идет. Мы также не в полной мере можем говорить о стратегии переадресации, поскольку изначальной адресации при создании текста машиной тоже не бывает. Алгоритм перевода АСТ ближе всего к стратегии терциарного перевода.

В рамках подстратегии терциарно-модерационного перевода при переводе АСТ должна учитываться необходимость обязательной модерации текста переводчиком по таким критериям современного автоматически сгенерированного текста, как наличие фактических ошибок, правильность грамматики и орфографии, верно выбранные языковые средства, включая стилистику и тональность, точность распознавания контекста темы (включая социальный, этнический, религиозный, географический и т.д. контекст), этичность («уровень токсичности») текста, его безопасность в плане использования конфиденциальных данных.

Автоматически сгенерированные тексты на русском и английском языках включают лексические, грамматические, синтаксические, стилистические, фактические, этические и другие погрешности. В связи с этим подстратегия терциарно-модерационного перевода АСТ должна включать проверку и премодерацию (предредактирование) автоматически сгенерированного текста по всем этим критерияльным направлениям. Такая стратегия должна применяться и при создании перевода АСТ переводчиком-человеком, и при подготовке АСТ к машинному переводу, поскольку существуют определенные ограничения машинного перевода, чувствительные к погрешностям текстов, а автоматическая генерация эти погрешности не исключает, а зачастую тиражирует.

Литература

1. *Илюхин В. М.* Стратегии в синхронном переводе: на материале англо-русских и русско-английских комбинаций перевода: дис. ... канд. филол. наук. – М., 2000. – 280 с.
2. *Сдобников В. В.* Перевод и коммуникативная ситуация. – М.: Флинта: Наука, 2015. – 470 с.
3. *Андреева Г. М.* Социальная психология: учебник для высших учебных заведений. – М.: Аспект Пресс, 2001. – 284 с.
4. Налет. Распределение по годам (частота на миллион словоформ) в основном корпусе с 1739 по 2021 // Национальный корпус русского языка [Электронный ресурс]. – URL: <https://ruscorpora.ru/s/aM8jO> (дата обращения: 16.04.2023).

5. *Благовецкий А.* «Вы продоёте рыбов». Картинка с котами стала мемом года // Российская газета [Электронный ресурс]. – URL: <https://rg.ru/2021/08/02/vy-prodoyote-rybov-kartinka-s-kotami-stala-memom-goda.html> (дата обращения: 16.04.2023).
6. *Floridi L., Chiriatti M.* GPT-3: Its Nature, Scope, Limits, and Consequences. *Minds & Machines* 30, 681-694 (2020) [Электронный ресурс]. – URL: <https://doi.org/10.1007/s11023-020-09548-1> (дата обращения: 16.04.2023).
7. *Kumar E.* *Natural Language Processing*. – International Pvt Ltd, 2011. – 202 p.
8. *McDonald D., Bolc L.* *Natural language generation systems*. – New York: Springer-Verlag, 1988. – P. 7-11.