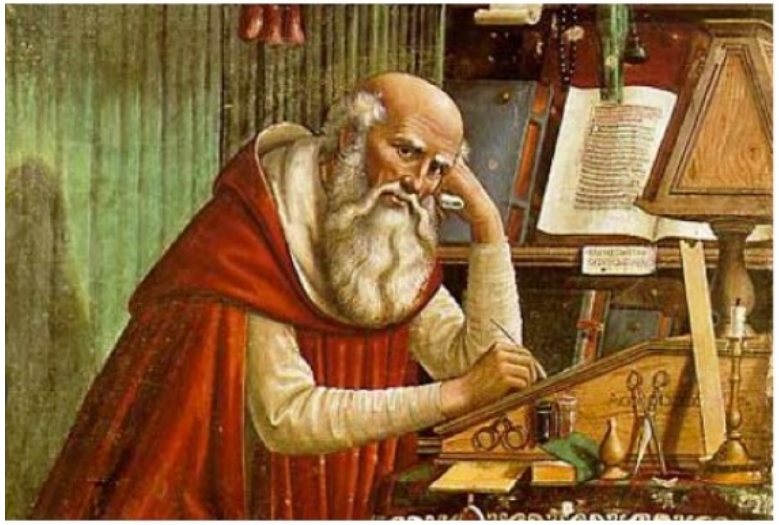
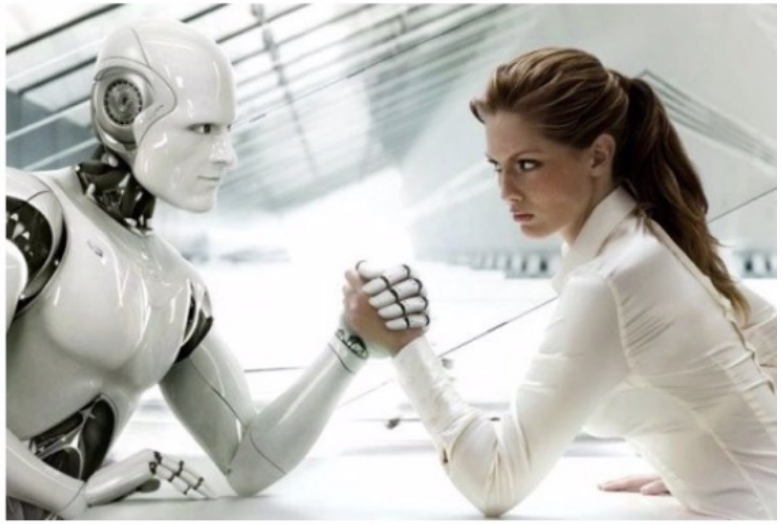
Постредактирование машинного перевода

КАФЕДРА ПЕРЕВОДА И ПЕРЕВОДОВЕДЕНИЯ ИГУ

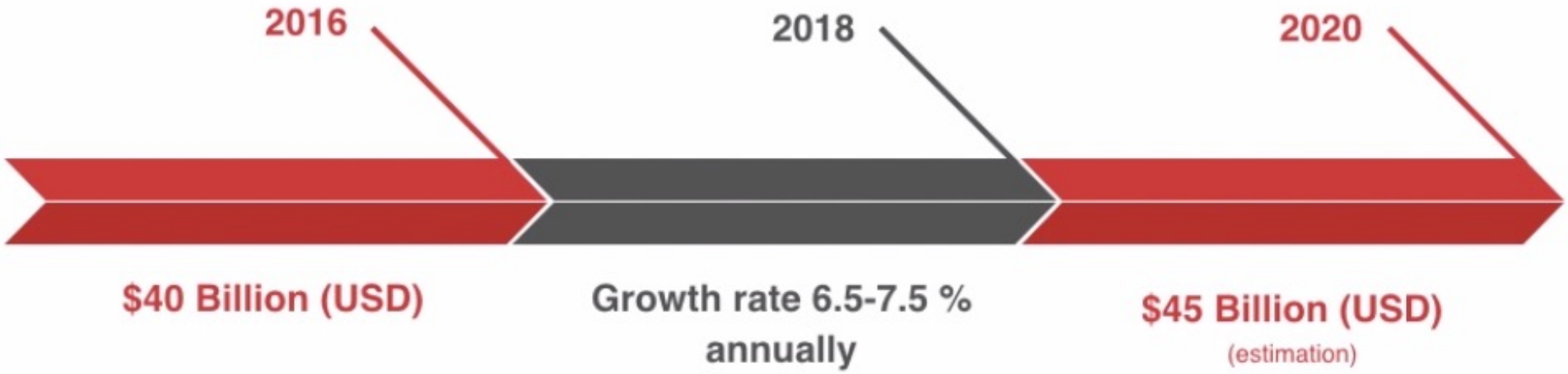
Память переводов vs. Машинный перевод





Навыки, требующиеся переводчику, изменились под запросы индустрии.
Переводчикам жизненно необходимо владеть ИТ-технологиями!

Value of Global Translation Industry



Language Technology is estimated at €29 Billion

Подходы к редактированию ПТ

Традиционный подход (полное редактирование)	**Постредактирование МП (когда имеется в виду запрашиваемый заказчиком вид деятельности)**
- исправление смысловых искажений и неточностей и проч., грамматических, орфографических, пунктуационных ошибок - один ИТ (исходный текст), - фокус на предложение (как единицу перевода) и уровень дискурса, - выбор стратегии перевода исходя из собственных предпочтений и требований к качеству текста. - Терминология точная, единообразная - Стиль и тон в соответствии с ИТ - Форматирование согласно оригиналу	- помимо ИТ имеется еще один текст – результат МП, - устранение вполне ожидаемых, часто встречающихся ошибок МП, - фокус на слова и фразы, - специфическая стратегия работы с текстом: минимум изменений, быстрая правка, - регулярный фидбек для улучшения качества МП - Если это легкое, быстрое, минимальное постредактирование, см. далее.

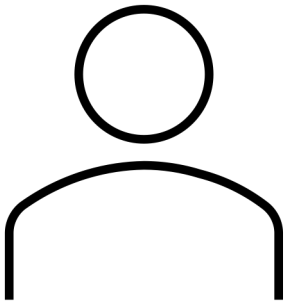
Легкое, минимальное, быстрое редактирование МП:

- Грамматические, пунктуационные, орфографические ошибки (в небольшом количестве) приемлемы.
- Лексика должна быть ясна, понятна, используема в данной сфере.
- В стиле и тоне изложения не должно быть ничего оскорбительного, агрессивного, остальное – допустимо.
- Стилистические погрешности, неточный подбор лексической единицы допустимы.
- Форматирование не в приоритете.

Требования и возможности редакторов МП

Способен выполнить: 4000-8000 слов в день

Скорость работы выше обычной в 2-3 раза



Это альтернатива более трудоемкому «человеческому» переводу

Спрос на редакторов МП растет

Постредактирование МП

- процесс обработки машинного перевода для достижения необходимого уровня качества для выполнения определенной задачи.

Выделяют два вида постредактирования – легкое и полное.

- При легком постредактировании допустимо наличие стилистических ошибок, терминология предлагается компьютерной программой.
- Единственный критерий к качеству – возможность понять текст без обращения к оригиналу (главным образом, тем, кто не владеет ИЯ).
- В результате полного постредактирования должен получиться стилистически соответствующий исходному понятный текст, точное и законченное сообщение, не отличающееся от сделанного человеком.

Еще один взгляд на варианты постредактирования МП (платформа TAUS)

- Различают «входящий» перевод (inbound translation), у которого имеется единственная цель – чтобы при прочтении был понятен в общих чертах смысл ИТ.

Редактирование МП человеком может вообще не выполняться, за исключением совсем уж одиозных случаев. Либо выполняется легкая правка МП без контроля качества перевода.

- «Исходящий» перевод (outbound translation) - редактирование МП с надлежащим качеством текста для публикации.

В этом случае иногда бывает достаточно минимального редактирования МП (такой тип редактирования используется, например, для перевода текстов тяжелого машиностроения и еще некоторых отраслей). В остальных случаях применяется полное постредактирование МП, при этом переводчик-редактор может пожертвовать стилистическим аспектом.

Критерии выбора стратегии

- знание предметной области,
- степень владения языками,
- степень доверия и опыт работы с данной системой МП,
- сроки и предполагаемый объем постредактирования и т.д.

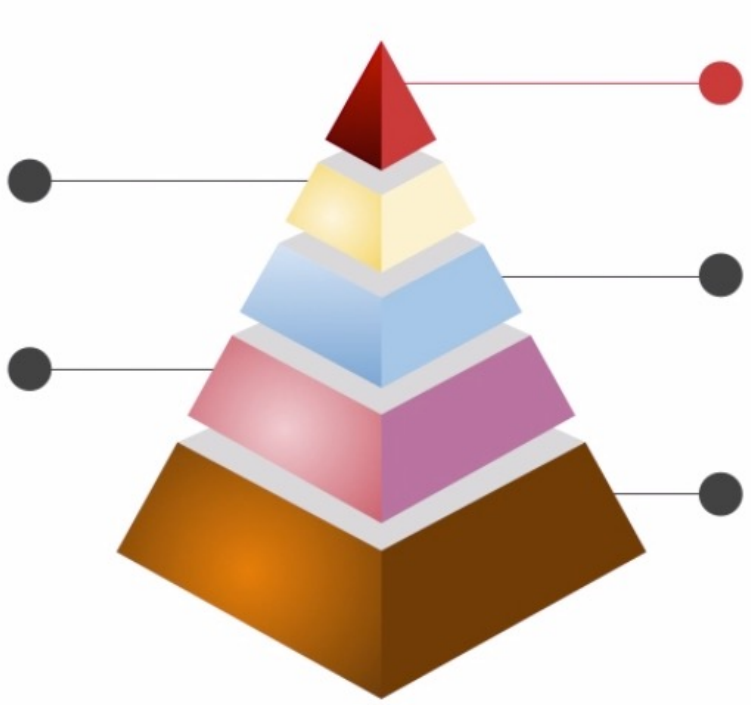
Стратегии, применяемые в полном постредактировании машинного перевода

- Полное редактирование выполняется с проверкой каждого компонента текста по исходному фрагменту;
- Чтение МП и исправление вызывающих сомнения мест с обращением к исходному тексту;
- Использование МП исключительно в качестве подсказки по терминологии.

Базовые принципы постредактирования МП (платформа TAUS)

Информация ИТ и ПТ должна совпадать полностью, без добавлений

Грамматические ошибки должны быть исправлены, однако на первом месте – передача смысла.



Где только возможно, необходимо использовать «сырой», без правок МП

Никаких правок стиля для большей «гладкости» изложения

Элиминация лексики, неприемлемой для культуры ПЯ

Основной принцип постредактирования МП – наличие подробных инструкций заказчика

Стратегии быстрого, легкого выполнения постредактирования МП:

- ИТ должен полностью отвечать требованиям, предъявляемым контролируемым естественным языкам. Это редко соблюдается на постоянной основе.
- Спец.терминология должна быть предоставлена в формате базы данных.
- ПТ должен быть понятно изложен, легко читаться.
- Вся информация без купюр излагается в ПТ без добавлений со стороны переводчика, даже для большей гладкости прочтения, т.е. никаких описательных переводов, добавлений, комментариев, пояснений, конкретизации и т.д.

- В приоритете – смысл, структура предложения, термины.
- Стиль редактировать необязательно, однако неприемлемая для культуры ПЯ лексика должна быть заменена или удалена.
- **Контроль качества должен включать в себя:**

- Исправление нарушений норм ПЯ;
- Неверно подобранные термины должны быть заменены на эквиваленты
- Даты, единицы измерения, и т.п. должны быть скорректированы под требования ПЯ. Теги CAT-программ и сегменты ИТ и ПТ должны соответствовать друг другу.
- Проверка орфографии и пунктуации.

- Соблюдение надлежащего форматирования необходимо.

Правила постредактирования МП

У бюро переводов имеются свои инструкции по выполнению разного рода постредактирования МП. Определяющими факторами являются система МП, тип текста, область знаний, отрасль и т.д., языковая пара и т.д.

трудно сформулировать универсальные правила, подходящие любой ситуации.

НО: в любом случае любой текст МП, подвергнутый постредактированию, должен пройти процедуру проверки качества, рамки которого устанавливает заказчик.

Что нужно проверить перед МП:

- Проверить, готов ли ИТ к машинному переводу;
- Проверить, какой тип МП будет использоваться, т.к. разного рода движки МП требуют дифференцированного подхода;
- Определить тип постредактирования МП, посоветовавшись с заказчиком (зависит об объема, бюджета, времени и прочих факторов);
- Установить уровень качества, который подходит заказчику (если необходимо, приведите примеры);
- Оценить качество «сырого» машинного перевода (без каких-либо правок со стороны человека);
- Проверить, был ли применен при МП интегрированный в движок терминологический модуль, подключен спец.словарь терминов и т.п.

Правила постредактирования МП

- Редактирование не должно быть равно по качеству, трудоемкости традиционному переводу текста с нуля.
- Действует правило двух секунд, т.е. все решения по редактированию должны приниматься быстро – нужно ли редактировать данный сегмент либо его нужно переводить самостоятельно.
- Непереведенные сегменты в тексте МП должны быть переведены либо классифицированы как не требующие перевода: названия фирм, изданий и т.д.
- Если в тексте МП есть предложения с неясным смыслом, они должны быть скорректированы, иначе ПТ будет непонятен получателю перевода.
- Неточно переведенные, бессмысленные предложения следует удалить, если их сложно довести до ума с помощью редактуры и проще было бы переводить с нуля самостоятельно.
- Значения лексических единиц и синтаксические структуры в приоритете.
- Повторы слов не следует заменять синонимами (разве что имеются специальные глоссарии или инструкции на этот счет).

Компетенции редакторов МП

- Знание ИЯ и ПЯ не ниже уровня С1;
- знание особенностей различных типов текстов и жанров, знание лингвистики текста (феномен текстуальности, процессы порождения текстов и функционирования высказываний разного речежанрового формата, когезия и когерентность и т.д.);
- Навыки работы с текстами;
- Навыки работы с глоссариями;
- Знание предметной области и навыки работы с текстами, относящимся к этой области;
- Знание технологий МП.

Как развить нужные навыки?

- выполнять постредактирование текстов МП минимум в течение месяца. Если говорить об объеме, то отредактировать минимум 100 000 слов.
- Интересоваться новыми технологиями, осваивать их;
- Выполнять постредактирование МП с помощью разных систем МП;
- Изучать настройки систем МП, научиться настраивать их, по возможности – отправлять фидбек разработчикам.;
- Пользоваться специальным ПО для управления и использования глоссариев и онлайн-словарями.
- Пользоваться при необходимости программами анализа и лингвистической обработки текстов (ссылка на такое ПО: https://rvb.ru/soft/catalogue/c01.html https://vallaam.livejournal.com/68186.html)

Процесс постредактирования (1)

- Просмотр текста для понимания его содержания в общих чертах;
- Обнаружение и систематизация гипотетических повторяющихся ошибок МП;
- Идентификация и перевод непереведенных сегментов текста МП;
- Обнаружение смысловых искажений и неточностей в тексте МП, устранение этих искажений;
- Удаление неадекватного перевода сегментов, если корректировка будет слишком трудоемкой, а смысл данных сегментов не является достаточно важным для таких временных затрат;
- Исправление синтаксических и лексических ошибок;
- Исправление стилистических ошибок;
- Удаление или замена неприемлемой лексики для культуры ПЯ;

Процесс постредактирование (2)

- Проверка сохранения единообразия терминологии;
- Проверка тегов;
- Проверка передачи цифровой информации и наличия двойных пробелов;
- Проверка терминологии;
- Проверка орфографии;
- Проверка форматирования.

Ключевые положения эффективного постредактирования МП (1)

- Использование списков терминологии, глоссариев (как заранее утверждённых и подготовленных (если возможно, то загруженных в системы МП заранее), так и пополняемых во время работы над ПТ);
- Знание особенностей работы выбранного типа движка МП (порождают ошибки разного рода);
- Выбор CAT-систем (для автоматизированного перевода) обуславливается наличием необходимых модулей: система управления глоссарием, наличие ТМ (память переводов), могут экспортировать файлы, подключать системы МП (по API) либо имеются свои системы МП;
- Быстрое предредактирование МП (возможные повторяющиеся ошибки, которые могут допускать определенные системы МП);
- Особое внимание на окончания лексических единиц в предложении;

Ключевые положения эффективного постредактирования МП (2)

- Новички имеют тенденцию переусердствовать с постредактированием, доводя перевод до совершенства;
- Задача постредактора МП – установить приемлемый уровень качества текста МП, редактировать только то, что необходимо;
- НО! Типичной ошибкой является обратная ситуация – когда без исправления остаются серьезные ошибки МП, мешающие пониманию содержания текста. Такие ошибки должны быть исправлены и на уровне предложения, и на уровне текста.
- Необходимо составить список терминов, не подлежащих переводу - они должны остаться на ИЯ.

Использование стандартных форматов в работе

Translation Memory Export format (TMX) and

Localization Interchange File Format (XLIFF),

Кроме этого, исходный текст должен быть чистым: никаких рисунков, графиков, тегов, артефактов, оставшихся от сканирования и проч. Чем чище текст, тем лучше будет результат МП.

Виды систем МП в зависимости от формата использования

Десктоп-решения (системы МП в виде программ находятся на серверах или компьютерах разработчиков), все необходимые файлы (ИТ), базы данных, глоссарии, правила и списки можно загружать через такие программы, которые подключаются к серверам разработчика МП напрямую.

Сервис-решения: не требуют установки специального ПО, являются в основном «облачными». Все файлы требуется загружать туда, по одному или пакетом.

Посредством плагинов или API системы МП можно установить также в практически любой CAT-инструмент (требует особых тех.компетенций).

Контроль качества постредактирования МП

- Необходимый шаг в деле постредактирования МП;
- Проверка форматирования необходима;
- Можно составить списки вопросов, что следует проверить, если это необходимо;
- Для проектов объемом 10 тыс. слов и больше – проверка качества проводится выборочно:

- ок. 10 тыс. слов: 2-3 предложения на каждые 200-300 слов;
- Больше 10 тыс. слов: следует выбрать ключевой отрывок текста в 3000 слов и проверить качество редактирования МП.

Обязательное применение всех средств автоматического контроля качества перевода, например, встроенного в CAT-платформы.

Пример (взят из статьи Абросимовой Н.А., Щелоковой Е.А. «Пред- и постредактирование МП медицинских текстов»

Оригинал	*Предредактирование (контролируемый текст)*
As *the HAS-BLED score* does not account for such *asymptomatic MRI lesions*, *the score* may *underestimate bleeding risk* in this *population* [10]	*The "HAS-BLED" score* may *underestimate bleeding risk* in this *population*. As *the "HAS-BLED" score* does not account for such *asymptomatic MRI lesions*

Используемые системы МП: DeepL, GNMT, Яндекс.Переводчик

DeepL Переводчик	*GNMT*	*Яндекс.Переводчик*
Показатель HAS-BLED может *недооценивать риск кровотечения* в этой *популяции*. Поскольку *показатель HAS-BLED* не учитывает такие *бессимптомные поражения при МРТ*	*Оценка HAS-BLED* может *недооценивать риск кровотечения* в этой *популяции*. Поскольку *оценка HAS-BLED* не учитывает такие *бессимптомные поражения на МРТ*	*Показатель HAS-BLED* может *недооценивать риск кровотечения* в этой *популяции*. Поскольку *оценка HAS-BLED* не учитывает такие *бессимптомные МРТ-поражения*

Оригинал	*Предредактирование (контролируемый текст)*
In these circumstances, we analysed the *MRI*-confirmed and *TOAST* only (*standard phenotyping*) groups separately and divided the *study controls* between the two groups to avoid any *sample overlap.* [11]	Therefore, the *MRI*-confirmed and "*TOAST*" only (*standard phenotyping*) groups were analysed separately. The *study controls* were divided between the two groups to avoid any *sample overlap*

DeepL Переводчик	*GNMT*	*Яндекс.Переводчик*
Поэтому группы, подтвержденные с помощью *MPT* и только *TOAST* (*стандартное фенотипирование*), были проанализированы отдельно. *Контрольные группы* исследования были разделены между двумя группами во избежание *дублирования выборки*	Поэтому группы с подтвержденным *MPT* и только *TOAST* (*стандартное фенотипирование*) анализировали отдельно. *Контрольная группа* исследования была разделена между двумя группами, чтобы избежать *дублирования выборки*	Поэтому группы, подтвержденные *MPT* и только для *ТОСТОВ* (*стандартное фенотипирование*), были проанализированы отдельно. *Контрольные группы* исследования были разделены между двумя группами, чтобы избежать какого-либо *перекрытия выборки*

Без предредактирования МП

Оригинал	*Deep Переводчик*	*GNMT*	*Яндекс.Переводчик*	*Постредактура*
Both *CAA* and *atrial fibrillation* increase significantly with age, and people may undergo brain *MRIs* for a variety of indications; therefore, *clinicians* are more commonly encountering patients with both *atrial fibrillation* and *CAA*, resulting in *clinical management dilemmas* [10]	Как *CAA*, так и *фибрилляция предсердий* значительно увеличиваются с возрастом, а люди могут проходить *МРТ* головного мозга по разным показаниям; поэтому *клиницисты* все чаще сталкиваются с пациентами как с *фибрилляцией предсердий*, так и с *CAA*, что приводит к *дилеммам клинического ведения*	Как *ВГА*, так и *фибрилляция предсердий* значительно увеличиваются с возрастом, и люди могут проходить *МРТ* головного мозга по разным показаниям; поэтому *клиницисты* чаще сталкиваются с пациентами как с *фибрилляцией предсердий*, так и с *ВГА*, что приводит к *дилеммам клинического ведения*	Как *CAA*, так и *фибрилляция предсердий* значительно увеличиваются с возрастом, и люди могут проходить *МРТ* головного мозга по различным показаниям; поэтому *клиницисты* чаще сталкиваются с пациентами как с *фибрилляцией предсердий*, так и с *CAA*, что приводит к *дилеммам клинического управления*	Пациенты могут проходить *МРТ-обследование* головного мозга по разным медицинским показаниям в связи с тем, что с возрастом возникновение риска *ЦАА* и *фибрилляции предсердий* существенно увеличивается. Врачи все чаще сталкиваются с *проблемами при оказании клинической помощи* пациентам как с *фибрилляцией предсердий*, так и с *ЦАА*

Оригинал	*DeepL Переводчик*	*GNMT*	*Яндекс.Переводчик*	*Постредактура*
Any cases with subsequent *MRI* confirmation of *lacunar stroke* were omitted from *the TOAST only group*; all *individuals* were analysed only once [11]	Все случаи с последующим *МРТ*-подтверждением *лакунарного инсульта* были исключены из *группы TOAST only*; все *люди* анализировались только один раз	Любые случаи с последующим *МРТ*-подтверждением *лакунарного инсульта* были исключены из *группы только TOAST*; все *особи* были проанализированы только один раз	Все случаи с последующим *МРТ*-подтверждением *лакунарного инсульта* были исключены из *группы «Только ТОСТ»*; все *пациенты* были проанализированы только один раз	Все *контрольные образцы* случаев с последующим подтверждением *лакунарного инсульта* только при помощи *МРТ* были исключены из *группы TOAST*, и проанализированы один раз

Выводы по данному исследованию (1)

- В большинстве случаев предредактирование положительно влияло на качество машинного перевода благодаря применению определенного алгоритма действий.
- При разработке контролируемых текстов стоит уделять особое внимание синтаксису, поскольку из-за сохранения структуры предложения на ИЯ системы машинного перевода нарушают логические связи между элементами предложения, из-за чего при передаче терминов отсутствует единство.
- В случае с медицинскими терминами предредактирование осуществить не представляется возможным, поскольку данные единицы изменению не подлежат, следовательно, существует вероятность возникновения ошибок при переводе медицинских терминов. Случаи, где термины даны с расшифровкой, являются исключением.

Выводы по данному исследованию (2)

- К машинному переводу медицинских текстов необходимо применять полное постредактирование в силу специфики данного типа текстов.

Нами применялись следующие стратегии:

- чтение машинного перевода и исправление вызывающих сомнения мест с обращением к исходному тексту;
- верификация терминов, даже если системой они были правильно идентифицированы и переведены, поскольку системы МП не всегда учитывают тематику текста и подбирают соответствия из других областей;
- проверка каждого компонента текста по исходному фрагменту;
- исправление синтаксиса, т. к. системы МП оставляют синтаксис ИЯ, что нарушает смысл и является недопустимым при переводе медицинского текста. Было установлено, что анализируемые в работе системы подходят только для полного постредактирования, поскольку необходимо редактирование большого объема информации для достижения приемлемого качества перевода, чтобы обеспечить читающему должное восприятие и понимание текста.

Вопросы?