И.Д. МАМАЕВ, Д.М. ШАМОВА

ПРИЛОЖЕНИЯ ДЛЯ АВТОМАТИЗАЦИИ ПЕРЕВОДЧЕСКОГО ПРОЦЕССА

Учебное пособие

Санкт-Петербург
Издательство БГТУ «ВОЕНМЕХ» им. Д.Ф. Устинова
2024

УДК 81’322.4(075.8)
М22

Мамаев, И.Д.

М22 Приложения для автоматизации переводческого процесса: учебное пособие / И.Д. Мамаев, Д.М. Шамова. – Санкт-Петербург: Изд-во БГТУ «ВОЕНМЕХ» им. Д.Ф. Устинова, 2024. – 35 с.

ISBN 978-5-00221-063-3

Рассмотрены технологии машинного и автоматизированного перевода, приложения для создания субтитров, а также ряд систем автоматической обработки текстов.

Предназначено для студентов, обучающихся по направлениям подготовки «Теоретическая и прикладная лингвистика», «Перевод и переводоведение», «Интеллектуальные системы в гуманитарной среде».

УДК 81’322.4(075.8)

Р е ц е н з е н т и.о. зав. кафедрой общего языкознания им. Л.А. Вербицкой СПбГУ, канд. филол. наук, доц. *Е.И. Риехакайнен*

Утверждено
редакционно-издательским
советом университета

ISBN 978-5-00221-063-3

1. СОВРЕМЕННЫЙ ПЕРЕВОДЧИК И ИНФОРМАЦИОННО-КОММУНИКАЦИОННАЯ СРЕДА. АВТОМАТИЗИРОВАННЫЙ ПРЕДПЕРЕВОДЧЕСКИЙ АНАЛИЗ

С появлением информационно-коммуникационных технологий (ИКТ) понятие перевода было переосмыслено, что вызвало существенный перелом не только в переводческой практике, но и в общей теории перевода. Изменения в традиционной переводческой деятельности связаны с несколькими факторами. Во-первых, специалисты в области компьютерной лингвистики разрабатывают огромное количество систем ***автоматической обработки текста (АОТ)***, которые позволили в разы уменьшить временны́е затраты на «производственный» процесс перевода. Во-вторых, существенную роль также сыграла «компьютеризация» учреждений, организаций, частного бизнеса, профессиональных рабочих мест и т.д. Компьютеры и приложения уже некоторое время являются неотъемлемой частью рабочего процесса переводчика. Наконец, разнообразие форматов электронных документов, а также скорость их распространения в Интернете привели к возникновению новой специализированной области – локализации, т.е. переводу интерфейсов, файлов компьютерных приложений, веб-сайтов. В результате современный переводчик должен иметь обширные знания в области компьютерных наук, которыми раньше обладали только специалисты.

А.А. Рыбакова утверждает, что профессиональная компетенция современного переводчика должна учитывать новые условия деятельности (рис. 1.1).

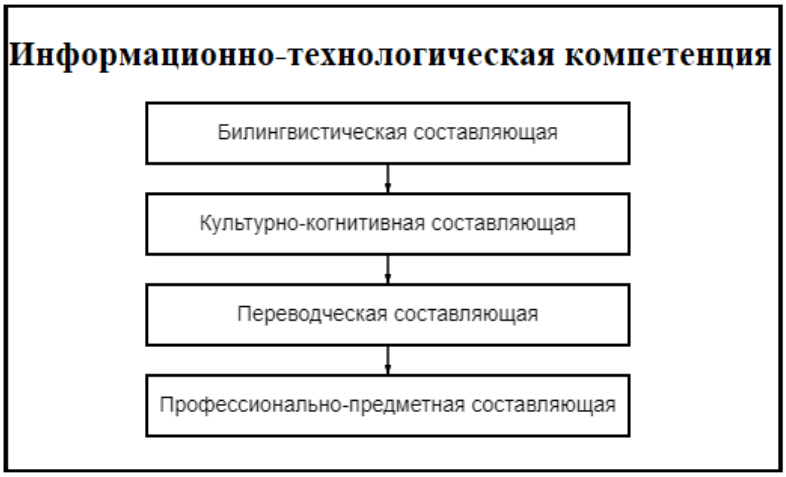

Рис. 1.1. Профессиональная компетенция современного переводчика

Компьютеры в новых информационных технологиях в переводе выступают как вспомогательный инструмент для специализированного программного обеспечения (ПО), обработки и хранения текстового материала. К основному переводческому ПО можно отнести электронные лексикографические ресурсы, размеченные корпуса текстов, системы автоматизации переводческого процесса и машинного перевода, текстовые редакторы, системы синтеза и распознавания речи, информационно-поисковые системы и пр.

В современном переводоведении успешная работа во многом зависит от выполненного предпереводческого анализа. За последние десятилетия были предложены многочисленные схемы анализа (например, схема И.С. Алексеевой или Кристианы Норд), которые имеют много общих черт. Традиционная процедура анализа заключается:

1) в сборе внешней информации о тексте;
2) определении целевой аудитории;
3) анализе плотности и состава информации;
4) определении коммуникативного задания;
5) определении речевого жанра.

Каждый из шагов так или иначе связан с ручной обработкой исходной информации: например, при анализе плотности и состава информации переводчик должен обращать внимание на сложность синтаксической структуры, основные тематики текста, наличие узкоспециализированных терминов и пр.

И.С. Алексеева в учебном пособии «Введение и переводоведение» обращает внимание на то, что опытному переводчику требуется в среднем 5-10 минут на анализ текста. Важно отметить, что при этом не оговариваются объемы текстовых массивов: с их увеличением будет расти и затрачиваемое время. На сегодняшний день в сети «Интернет» переводчики могут найти платформы, которые способны быстро обработать тексты и выдать необходимую для анализа информацию. Условно существующие платформы делятся на три группы:

1) code applications – кодовые приложения, требующие знания некоторых языков программирования (Python и пр.);

2) low-code applications – приложения, в которых пользователю, несмотря на автоматизацию большей части функций, требуется знать основы программирования (например, для написания простейших скриптов);

3) no-code applications – приложения, в которых пользователь без навыков программирования может свободно работать с заранее автоматизированными функциями.

Для переводчика как специалиста гуманитарной сферы деятельности большой интерес представляют платформы второй и третьей групп, ведь именно они позволяют обойти техническую сторону процесса обработки текстов и получить репрезентативные данные. Рассмотрим некоторые из приложений.

Orange – это low-code приложение для визуального программирования, в котором пользователи в своей рабочей среде при помощи соединения определенных функций получают «конвейер» обработки текстовых данных. На рис. 1.2 представлен стандартный процесс обработки.

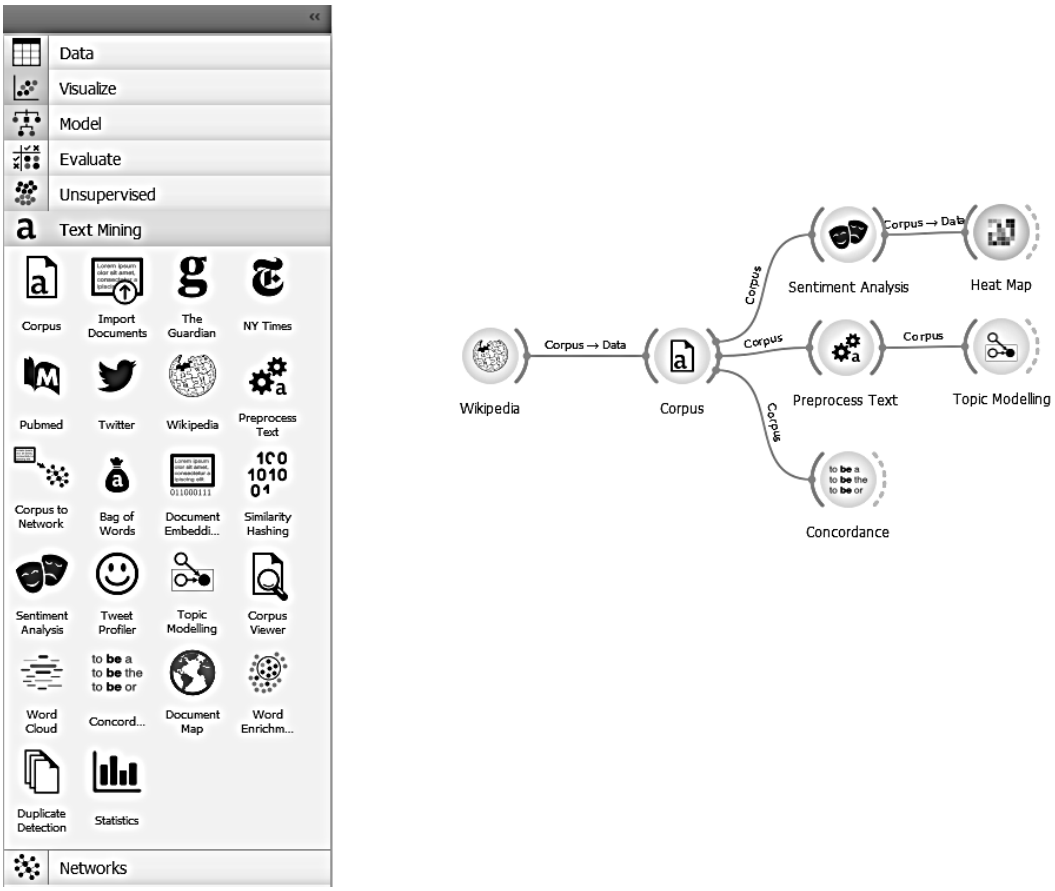

Рис. 1.2. Обработка статей из Википедии

На данном рисунке мы выбрали вкладку *Text Mining*, на которой пользователь может прибегнуть к одной из доступных функций. При работе с Orange переводчик может:

- провести тональный анализ исходного текста, что позволит подобрать соответствующие лексические единицы при переводе на целевой язык;
- определить тематические группы, которые помогут сформировать представление о целевой аудитории и речевом жанре;
- извлечь смысл лексемы с помощью автоматического построения конкорданса по тексту или коллекции текстов.

Функция *Wikipedia* позволяет отобрать необходимые статьи по заданным ключевым словам. В рамках данного примера были отобраны 22 англоязычные лингвистические статьи. С помощью *Corpus* пользователь сможет просмотреть текст статей, а потом приступить к анализу. Так, например, *Preprocess Text* обрабатывает текст: выделяет токены, удаляет стоп-слова и добавляет коллокации. В результате этих процедур пользователь может присоединить функцию *Topic Modelling* и узнать основные тематики в исследуемой выборке текстов.

В данной функции пользователь может выбрать один из трех алгоритмов тематического моделирования: Latent Semantic Indexing, Latent Dirichlet Allocation и Hierarchical Dirichlet Process. Например, в алгоритме Latent Dirichlet Allocation (Латентное размещение Дирихле) каждый документ – это набор тематик, с помощью которого на основе статистических вычислений можно выделить наиболее вероятные. Приведенные тематические множества показывают, что исследуемая выборка лингвистических статей посвящена грамматическим процессам, фонетике, словообразованию и пр. (рис. 1.3).

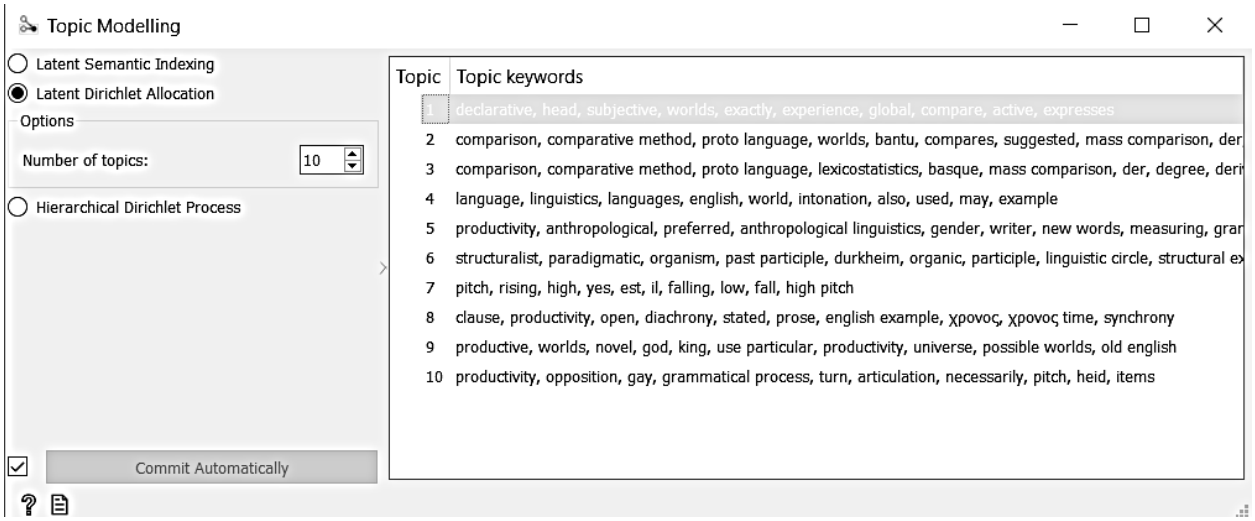

Рис. 1.3. Основные тематики лингвистических статей

Стоит также сказать и о функции *Sentiment Analysis*, в котором пользователь может выбрать один из существующих алгоритмов тонального анализа. Алгоритм *SentiArt* позволяет выделить базовые тональные аспекты в текстах художественной и научной литературы: злость, удивление, радость и пр. Функция *Heat Map* визуализирует результат (рис. 1.4).

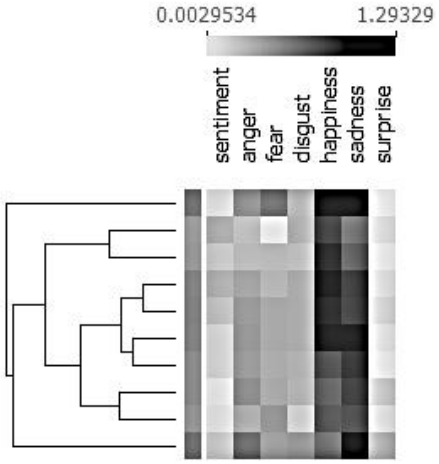

Рис. 1.4. Основные тональные компоненты

Таким образом, пользователь понимает, что преобладающие тональные оттенки – это *happiness* и *sadness*, т.е. исследуемые статьи являются нейтральными.

В рамках данного пособия невозможно рассмотреть весь функционал данного приложения, на официальном сайте можно ознакомиться с другими лингвистическими инструментами.

Stanza – это веб-приложение, которое также позволяет проанализировать текст на разных языковых уровнях: выявить морфологические характеристики лексических единиц, построить синтаксические деревья и пр. Для предпереводческого анализа рассмотрим функцию *Named Entities*, которая выделяет именные сущности. К именованным сущностям можно отнести имена людей, названия организаций и пр. При работе с именованными сущностями переводчик должен понять, какую переводческую тактику нужно выбрать: транскрипцию именованной сущности, транслитерацию или же сохранение названия на исходном языке.

Рассмотрим пример из учебника по переводу публицистики Д.М. Бузаджи: *…The inquest heard that **Joao Paulo Lusakumunu Kiese,** 38, a lay preacher from **Manor Park,** east **London,** visited **Dr Bernard Delvigne** at his private clinic in **Wimpole Street,** central **London,** with breathing problems last March…* В данном предложении мы вручную выделили несколько именованных сущностей: имена собственные и географические названия.

На рис. 1.5 к уже выделенным единицам были присоединены год и название месяца. Тем не менее можно сказать, что Stanza в разы быстрее справилось с определением сущностей (буквально за несколько секунд), поэтому пользователь уже без проблем сможет определить тактику перевода имен. В данном случае при передаче имен на русский язык лучший вариант – воспользоваться транскрипцией.

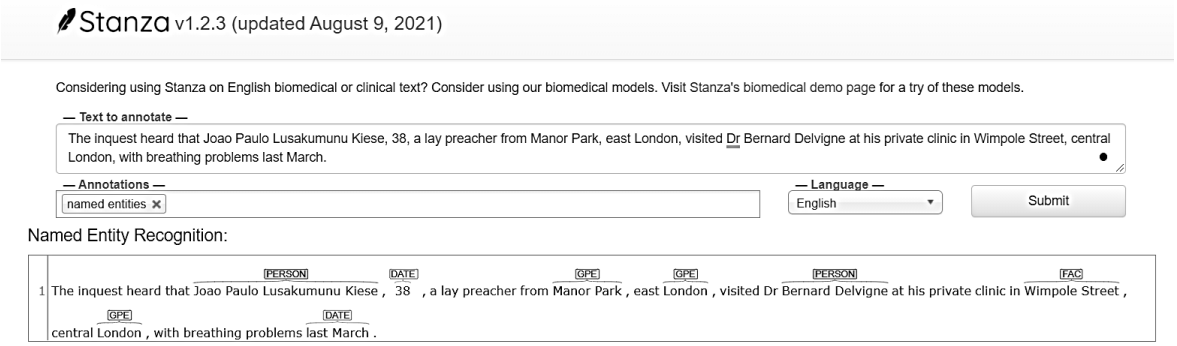

Рис. 1.5. Выделение именованных сущностей

Наконец, остановимся на приложении Voyant – веб-ресурсе, который позволяет без предварительной установки обработать текстовый массив документов. Введенный или загруженный текст будет подвержен лингвостатистическому анализу, в результате чего пользователю станут доступны частотные списки слов и коллокаций, «облако» терминов, встречаемых в текстах, графики распределения относительных частот слов в каждом отдельно взятом текстовом сегменте и пр.

На рис. 1.6 представлена основная страница, на которой пользователю предлагается ввести или загрузить свой текст. В данном случае остановимся на анализе лингвистической статьи. Кнопка *Reveal* предназначена для выдачи результатов (рис. 1.7).

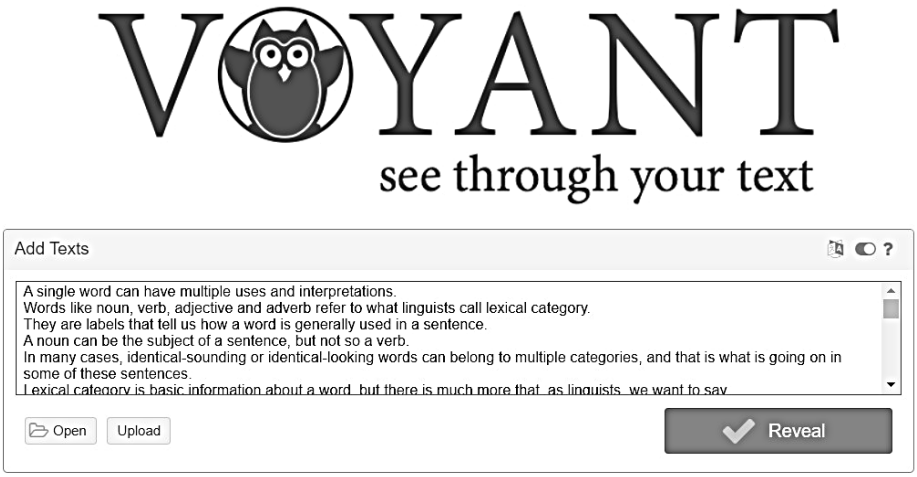

Рис. 1.6. Интерфейс начальной страницы

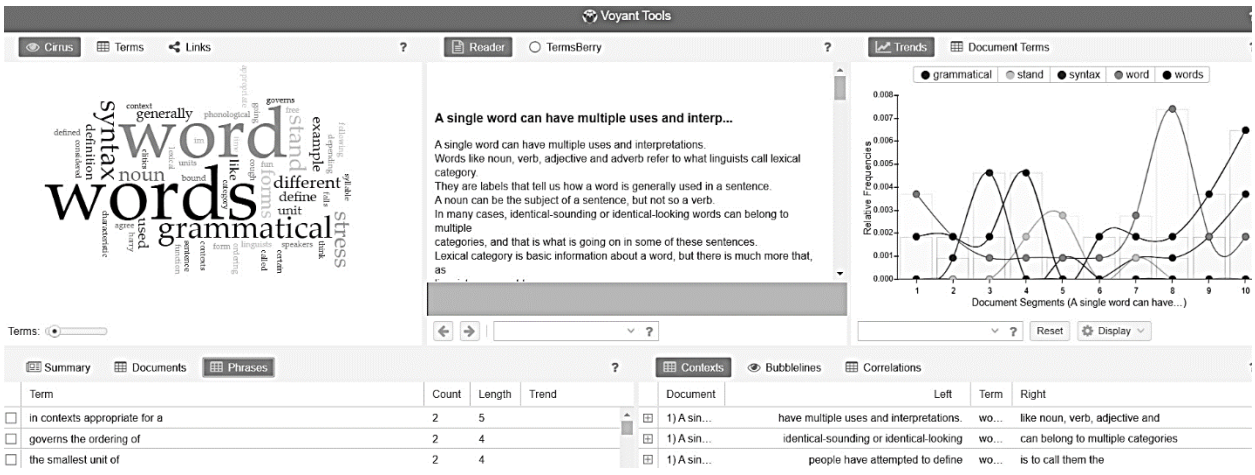

Рис. 1.7. Результат автоматического анализа

Пользователь даже при первом ознакомлении сразу сможет понять, что текст посвящен слову как лингвистическому понятию и его анализу с точки зрения грамматики и синтаксиса. Раздел *Phrases* поможет переводчику понять, какие выражения в целевом языке необходимо будет подобрать при переводе с исходного языка: например, для коллокации *the smallest unit of* можно подобрать эквивалент *наименьшая единица*. При переходе на вкладку *TermsBerry* пользователь ознакомится с необходимыми для перевода тематическими терминами.

При этом слова в исходном тексте не приводятся к канонической (словарной, начальной) форме, что приводит к неточному статистическому анализу. Уже на вкладке *Cirrus* видно, что *word* и *words* будут распознаны как две разные лексические единицы. Если лингвисту или переводчику необходимо провести строгий статистический анализ текста, то ему заранее придется привести к канонической форме все слова из текстового массива. Он может сделать это вручную или с помощью специальных приложений, известных как лемматизаторы.

2. ИНСТРУМЕНТАРИЙ АУДИОВИЗУАЛЬНОГО ПЕРЕВОДЧИКА. СОЗДАНИЕ СУБТИТРОВ

2.1. Основные принципы создания субтитров

Мир современных технологий требует современных решений. Технологический прогресс коснулся и сферы перевода: кинематограф вышел на международный уровень, игровая локализация все больше приобретает кинематографичный характер, инструменты для создания медиаматериалов становятся более доступными, популярность набирают как обучающие видеоролики, так и развлекательные видео: от кулинарных рецептов до выступлений комиков и прохождений игр.

Таким образом, возникает необходимость перевести все многообразие иноязычного контента и, что более важно, сделать это качественно. Для этого переводчику необходимо знать и учитывать специфику аудиовизуального перевода как особой деятельности, не относящейся ни к устному, ни к письменному переводу в полной мере. Если мы обратимся к английскому термину *audiovisual translation*, мы традиционно отметим, что *translation* – это перевод именно письменный, но окажемся правы лишь частично.

Аудиовизуальный перевод – перевод, ограничиваемый внеязыковыми рамками. Так как далее мы рассмотрим программы для субтитрирования, возьмем перевод под субтитры и изучим его подробнее. У субтитров почти всегда не более двух строк, они появляются и исчезают вместе с произносимой репликой. Иногда мы не успеваем читать субтитры: они слишком быстро пропадают или реплика очень длинная.

Субтитры не должны мешать зрителю, не должны разрушать иллюзию погружения в фильм. В данном разделе рассмотрена технология создания и редактирования двухмерных субтитров при помощи двух специализированных программ, доступных для скачивания бесплатно: ***Subtitle Edit*** и ***Aegisub***.

Программы функционируют по схожему принципу, но каждая из них имеет свои преимущества. Как показывает практика, для достижения высокого качества субтитров в некоторых случаях

необходимо применять несколько программ последовательно. Все в них устроено так, чтобы переводчик мог создать субтитры, которые соответствуют когнитивным возможностям зрителя. Для этого нужно соблюдать следующие правила:

1) субтитр не пересекает границу кадра, в котором находится;
2) субтитр не превышает определенное количество знаков в строке (***Characters per line*** или ***CPL***);
3) субтитр не превышает количество знаков в секунду (***Characters per second*** или ***CPS***).

Для стандартизации процесса создания субтитров компании-вещатели (*BBC, Disney, France 2, Netflix* и др.) выпустили специальные документы, называемые ***руководствами по стилю*** или *стайл-гайдами*. Ниже представлены выдержки из руководства по стилю компании «Нетфликс».

Руководство включает в себя общие требования к оформлению субтитров и требования для определенного языка. Рассмотрим требования, предъявляемые к русскоязычным субтитрам.

2.2. Требования к оформлению субтитров

Общие требования

1. Длительность:
   - минимальная длительность субтитра: 5/6 (пять шестых) секунды;
   - максимальная длительность: 7 секунд на каждый субтитр.
2. Формат файла: самый частотный формат субтитров – ***SubRip (*srt).***
3. Работа со строками:
   - максимальное количество строк – 2;
   - если количество символов в одной строке не превышает норму и вмещает необходимый текст, субтитр не разбивается;
   - при разделении субтитра на две строки действуют следующие правила:
     - разбить строку на две части можно: после знаков препинания; перед союзами; перед предлогами;
     - разрыв строки не должен отделять: существительное от артикля; существительное от прилагательного; имя от фамилии; глагол от подлежащего местоимения; предложный глагол от его предлога; глагол от вспомогательного глагола, возвратного местоимения или отрицания.
4. Расположение субтитров:
   - выравнивание субтитров – по центру;
   - стандартное расположение – внизу экрана, допустимо перемещение субтитров в верхнюю часть экрана, если фон внизу затрудняет восприятие текста или закрывает текст на экране, предусмотренный самим видеоматериалом (например, титры);
   - японский язык допускает вертикальное расположение субтитров.
5. Единообразие: таблицы ключевых имен и обращений должны быть созданы и использованы для перевода, чтобы обеспечить согласованность повторяющейся информации между эпизодами и сезонами.

Требования к субтитрам на русском языке

1. Сокращения: в аббревиатурах не должно быть точек: *ЦРУ, США, ООН.*
2. Ограничение по количеству символов: 39 символов в строке.
3. Имена персонажей:
   - имена собственные и прозвища транслитерируются;
   - прозвище переводится только в том случае, если оно передает свойство, важное для конкретного сюжета;
   - при необходимости перевода исторических деятелей, персонажей и мифических существ учитывается специфика языка перевода;
   - для транслитерации имен собственных следует обратиться к руководствам и ссылкам, приведенным в библиографическом списке.
4. Многоточие:
   - при включении многоточий в субтитры используется один символ, код «U+2026», а не три точки или точки с запятой подряд;
   - многоточие не используется, когда субтитр разделен на два без пауз, речь говорящего не прерывается:

Строка 1: *Я не сомневался,*
Строка 2: *что вы согласны со мной*;

• многоточие используется для обозначения паузы (две или более секунд) или в случае, если реплика резко прерывается;

• если после паузы высказывание продолжается в следующей строке, многоточие используется и в начале второй строки субтитра:

Субтитр 1: *Дайте мне подумать…*
Субтитр 2: *... возможно, вы можете сделать это по-другому.*
– Я давно хотел тебе сказать…
– Не надо, я ничего не хочу знать!

Здесь же представлено оформление субтитров с учетом разных говорящих: первую реплику произносит один персонаж, а вторую, в ответ, уже другой; многоточие используется без пробела, чтобы указать, что субтитр начинается с середины предложения: *...подписал соглашение*.

5. Шрифты:

• стиль шрифта: *Arial* или *SansSerif*;

• размер шрифта: зависит от размера экрана;

• цвет шрифта: белый.

6. Диалоги на иностранном языке:

• диалог на иностранном языке следует переводить только в том случае, если предполагалось, что зритель его поймет (т.е. если в оригинальной версии он был снабжен субтитрами);

• незнакомые иностранные слова и фразы следует либо перевести (если предполагается, что их нужно понять зрителю), либо оставить на языке оригинала, в зависимости от творческого замысла;

• если текст остается на языке оригинала, нужно выделить его курсивом;

• при использовании иностранных слов следует проверять орфографию, ударения и пунктуацию.

7. Курсив: курсивом выделяется:

• диалог, который слышится через электронные устройства: радио, телефон, компьютер, телевизор; курсив используется только в тех случаях, когда говорящий находится вне сцены, а не просто за кадром или вне камеры;

• текст песни (если на нее предоставлены права);

• голос за кадром;

• в названиях альбомов, книг, фильмов, телешоу, песен, видеоигр и т.д. не следует использовать курсив – вместо него применяются шевроны («»).

8. Числительные;

• числа от 1 до 10 во всех склонениях (всех падежных формах) пишутся прописью: *один, два, три – одного, двух, трех*;

• числа больше десяти пишутся цифрами: *11, 12, 13*;

• когда предложение начинается с числа, это число всегда пишется прописью.

9. Время: время следует указывать в 24-часовом формате.

10. Меры измерения:

• в качестве десятичного разделителя используется запятая: *23,99*;

• не используется пробел между числом и знаком процента, например: *25%*;

• при работе с процентом, выраженным однозначным числом, число записывается прописью в нужной падежной форме: *три процента, трехпроцентный* и т.д.

11. Кавычки:

• кавычки используются в начале цитаты и после последней строки цитаты, т.е. отмечается начало и конец цитаты, а не начало и конец каждого субтитра в цитате;

• для обычных цитат используются шевроны («») без пробелов; для кавычек внутри кавычек – двойные прямые кавычки ("");

• кавычки используются, когда видно, что персонаж читает вслух;

• если экранный персонаж при разговоре использует слово в переносном смысле, следует ставить кавычки перед эквивалентным словом на языке перевода, чтобы сохранить творческий замысел и обеспечить ясность в отношении того, к какому слову или части предложения относятся воздушные кавычки;

• используются шевроны («») без пробелов для названий литературных, музыкальных и художественных произведений, музыкальных групп, спортивных команд, телешоу.

12. Ограничения скорости чтения:
- видеоматериалы для взрослых: до 17 символов в секунду;
- видеоматериалы для детей: до 13 символов в секунду.

13. Повторы: слова или фразы, повторяемые более одного раза одним и тем же говорящим несколько раз подряд, переводятся один раз.

2.3. Обзор программ для создания субтитров

Subtitle Edit. В Subtitle Edit экран делится на три части: непосредственно редактор субтитров, справа от него расположен видеоплеер, а в нижней части экрана находится осциллограмма звуковой дорожки (рис. 2.1). На ней отражены колебания звука, по которым можно отметить начало и конец реплики. Вы можете выделять фрагменты на осциллограмме (они окрасятся красным) или перемещать границы выбранного фрагмента. Выше расположен непосредственно субтитр с основными его показателями: *total length* – количество знаков в субтитре (CPL) и *chars/sec* – количество знаков в секунду (CPS). В отличие от количества знаков в субтитре или строке субтитра, знаки в секунду вручную посчитать нельзя, этот показатель высчитывается автоматически.

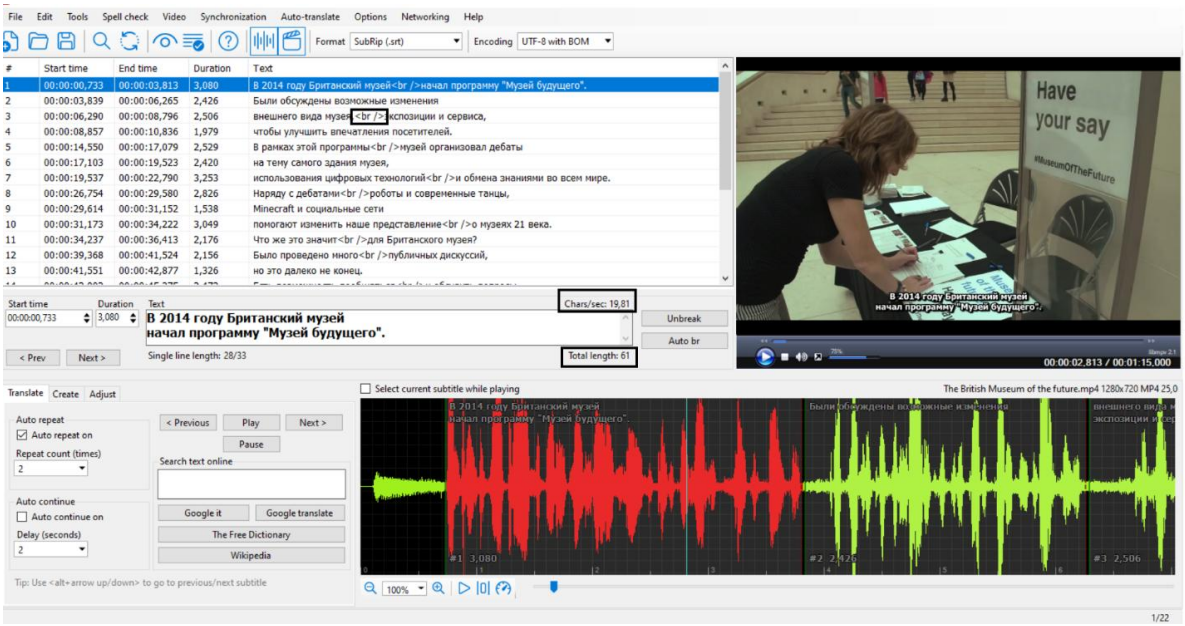

Рис. 2.1. Интерфейс программы *Subtitle Edit*

Для того чтобы создать субтитр, необходимо выделить фрагмент на осциллограмме, щелчком правой кнопки мыши на фрагменте вызвать контекстное меню и выбрать *Add text here* – в редакторе субтитров появятся соответствующие таймкоды: время начала и конца субтитра. Для создания субтитра из двух строк используется тег $< br/>$, при настройках по умолчанию тег ставится нажатием клавиши *Enter*, когда курсор находится в месте, где необходимо разбить субтитр.

Subtitle Edit позволяет настроить профиль пользователя. В профиле указываются границы показателей, при несоблюдении этих границ программа окрасит превышенный (или пониженный) показатель красным или выдаст сообщение о наползании субтитров друг на друга (*overlap*). Присутствуют и предустановленные профили, соответствующие руководствам по стилю различных компаний (рис. 2.2).

При редактировании субтитров необходимо их объединять и разбивать, переносить и удалять, все эти функции расположены в контекстном меню, вызываемом щелчком правой кнопки мыши на субтитре в редакторе.

Готовую работу можно сохранить во множестве форматов, однако самым распространенным остается формат ****srt (SubRip).***

Aegisub. Программа для субтитрирования Aegisub работает по схожему с Subtitle Edit принципу и так же имеет видеоплеер и аудиодорожку. Отображение определенных элементов настраивается поль-

зователем. К преимуществам данной программы можно отнести широкий функционал для цветовой разметки субтитров при помощи менеджера стилей, а также широкий набор дополнительных функций при работе с переводом субтитров.

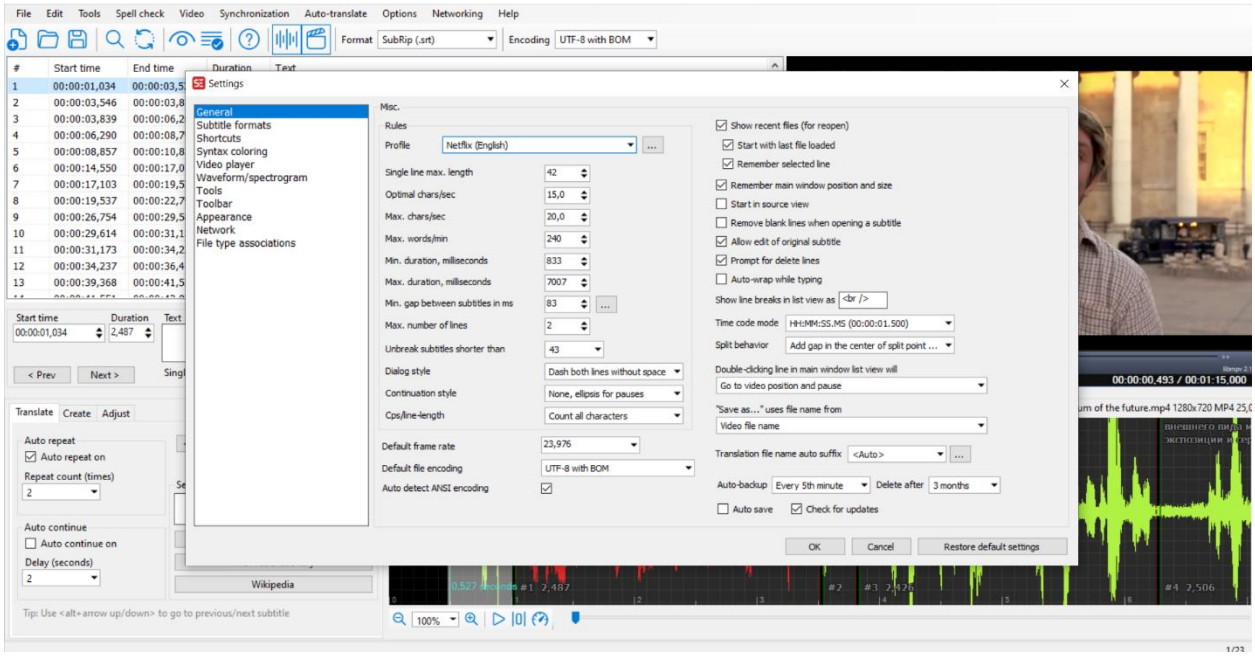

Рис. 2.2. Профиль пользователя в Subtitle Edit (*Options* → *Settings*)

Для начала работы необходимо загрузить видеоролик (рис. 2.3).

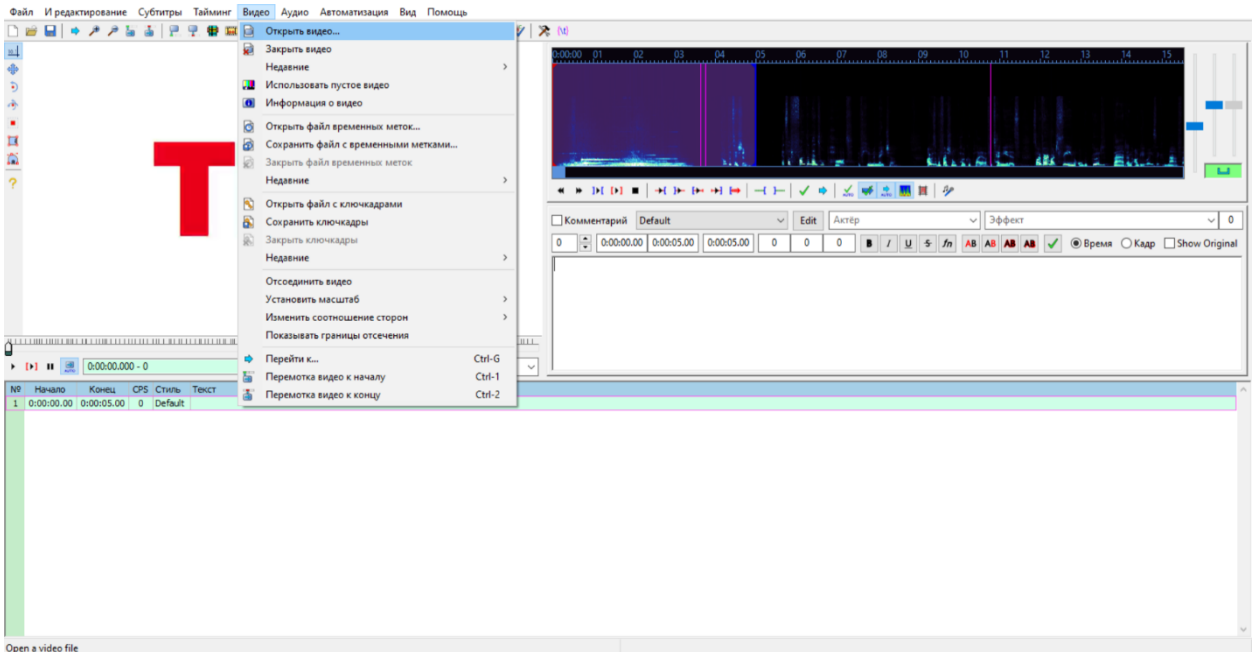

Рис. 2.3. Рабочее окно программы Aegisub

Загрузив видео, мы получим аудиодорожку, на которой, как и в Subtitle Edit, можно выбрать фрагмент и создать субтитр с соответствующей точкой входа и выхода (начала и конца субтитра) нажатием кнопки *Enter*. В окне отображения субтитров располагаются таймкоды, показатель скорости чтения (количество знаков в секунду или CPS), окрашиваемый красным при превышении, стиль строки субтитра и текст субтитра непосредственно.

Рассмотрим подробнее окно редактирования субтитров, а также перечислим основные функции, к которым может прибегнуть пользователь (рис. 2.4).

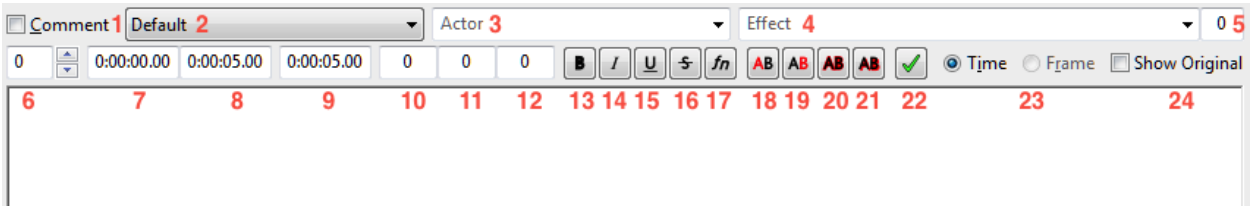

Рис. 2.4. Окно редактирования субтитров в программе Aegisub

1. Помечает строку как комментарий. Субтитры – комментарии не будут отображаться на видео.
2. Стиль, используемый для этого субтитра.
3. Актер, произносящий эту строку. Информация не отображается в субтитрах, но может оказаться полезной при редактировании.
4. Эффект для этого субтитра. Есть несколько предопределенных эффектов, которые могут применяться, но они не всегда поддерживаются плеерами. Поэтому для стилизации субтитров используются теги. Теги выступают в качестве метаданных для активации сценариев автоматизации.
5. Количество символов в самой длинной строке этого субтитра.
6. Слой субтитра. Если два субтитра отображаются друг на друге, можно выбрать, какой из субтитров будет отображаться поверх другого.
7. Время начала субтитра.
8. Время окончания субтитра.
9. Общая длительность субтитра.
10. Создание отступа слева.
11. Создание отступа справа.
12. Создание вертикального отступа.
13. Вставка тега для выделения субтитра жирным шрифтом ($\backslash b1$).
14. Вставка тега для выделения текста курсивом ($\backslash i1$).
15. Вставка тега для подчеркивания ($\backslash u1$).
16. Вставка тега для зачеркивания ($\backslash s1$).
17. Открытие окна выбора шрифта и стиля начертания букв ($\backslash fnFontName$) с заданным названием шрифта, а также теги уже примененных эффектов.
18. Открытие окна выбора цвета. Позволяет выбрать цвет, затем вставить тег основного цвета ($\backslash c$).
19. Открытие окна выбора цвета. Позволяет выбрать второй цвет, затем вставить тег дополнительного цвета ($\backslash 2c$). Программа позволяет использовать до четырех цветов.

Переход к следующей строке субтитра происходит при нажатии кнопки 22, при необходимости создается новая строка в конце файла. Обратите внимание, что в отличие от предыдущих версий Aegisub в последней на данный момент версии изменения не нужно подтверждать нажатием этой кнопки.

Рассмотрим контекстное меню программы Aegisub (рис. 2.5).

1. ***Средство проверки орфографии.*** Если вы щелкните правой кнопкой мыши на слове, написанном с ошибкой, программа проверки орфографии предложит статистически частотную замену. Вы можете указать, на каком языке будет проходить проверка, а также самостоятельно добавлять слова, чтобы исключить случаи распознавания орфографически верных слов как ошибок.
2. ***Тезаурус.*** Предлагает синонимы выделенного слова.
3. ***Разбиение субтитров.*** Позволяет разбить субтитр на два отельных субтитра там, где находится курсор. Функция *Preserve time* сохраняет таймкоды исходного субтитра для двух новых субтитров. Функция *Estimate times* создаст новые таймкоды, основываясь на длине выделенного курсором фрагмента.

На рис. 2.6 представлено контекстное меню и его основные функции.

В данном меню располагаются варианты объединения, присоединения, копирования и вставки субтитров в различных позициях: перед или после меток курсора, перед началом кадра или после его окончания.

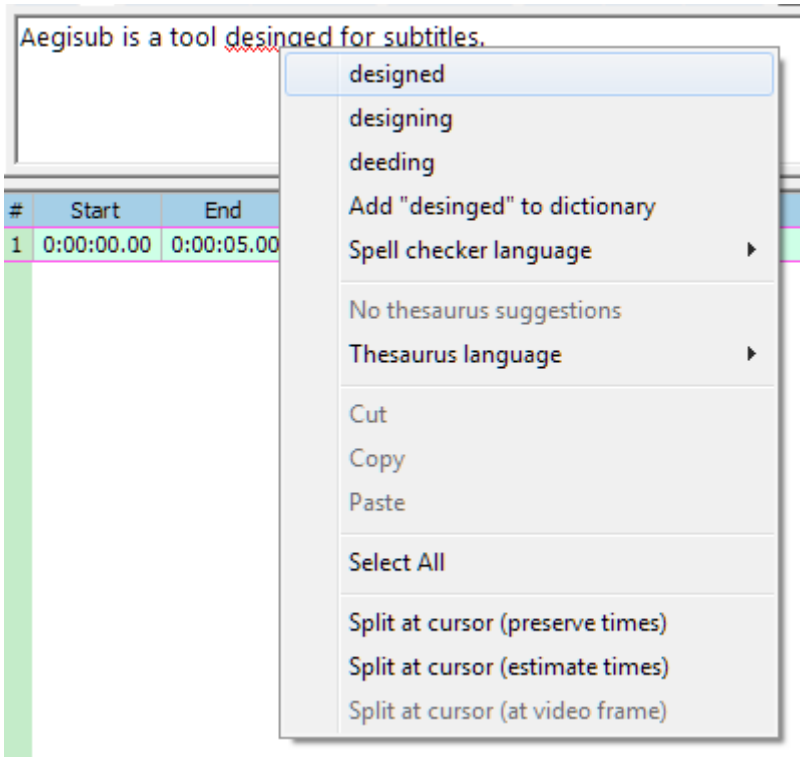

Рис. 2.5. Контекстное меню программы Aegisub

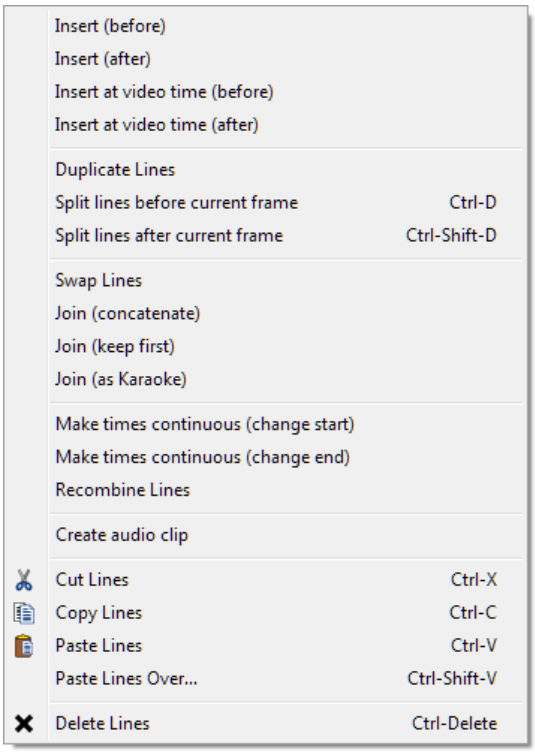

Рис. 2.6. Контекстное меню программы Aegisub, вызываемое щелчком правой кнопки мыши

Готовые субтитры пользователь может сохранить в формате *SubRip*, но рассматриваемая программа имеет и собственный формат – ***ass**. В отличие от известного формата *SubRip* (***srt**), формат субтитров *Aegisub* (***ass**) сохраняет и воспроизводит заданные для него стили: шрифт и цвета́.

3. МАШИННЫЙ ПЕРЕВОД И ПОСТРЕДАКТИРОВАНИЕ

Под ***машинным переводом,*** по мнению Уильяма Джона Хатчинса (*William John Hutchins*, 1939–2021), лингвиста и специалиста по информационным технологиям, понимается перевод с одного естественного языка (исходного языка) на другой язык (целевой язык) при помощи компьютерных систем и с помощью или без помощи человека (*…translation from one natural language (source language (SL)) to another language (target language (TL)) using computerized systems and, with or without human assistance…*). На протяжении многих лет машинный перевод привлекал лингвистов, специалистов по компьютерным наукам, философов и психологов. Первые работы в этой области оказали огромное влияние на развитие таких областей знаний, как компьютерная лингвистика и искусственный интеллект. В данном разделе подробно рассмотрены история машинного перевода, основные подходы к созданию данных систем и процесс постредактирования текстов с точки зрения современного переводоведения.

3.1. История машинного перевода

В XVII в. европейские философы предложили создать механизированные словари, чтобы упростить коммуникацию между иностранцами, однако этот подход так и остался на идейном уровне. Лишь в первой половине XX в. были предложены практические реализации. В 1924 г. в эстонской газете Vaba maa опубликовали новость о том, что изобретатель А. Вахер устроил демонстрацию собственной модели пишущей машины-переводчика. В 1933 г. независимо друг от друга такие системы были предложены Жоржем Арцруни (*George Artsouni*, жил и работал в первой половине XX в.), французским ученым армянского происхождения, и Петром Петровичем Смирновым-Троянским (1894–1950), советским инженером. Ж. Арцруни назвал свою систему «Механический мозг», в ее основе лежал большой механизированный словарь на бумажной ленте. П.П. Смирнов-Троянский описал «Машину для подбора и печатания слов при переводе с одного языка на другой или на несколько других одновременно», принцип которой можно представить следующим образом. Для корректной работы машины требуется два человека, первый из которых владеет исходным языком, второй – целевым. Сначала первый человек переводил все слова текста в каноническую форму, только после этого осуществлялся перевод в системе. Новый текст содержал множество слов на иностранном языке в канонической форме. На данном этапе второй человек приводил полученный текст в литературную форму, восстанавливая синтаксические связи и внося лексико-семантические изменения.

Использование компьютеров для переводческих целей приблизительно в то же самое время обсуждалось и за рубежом Уорреном Уивером (*Warren Weaver*, 1894–1978) и Эндрю Дональдом Бутом (*Andrew Donald Booth*, 1918–2009), специалистами по теории информации. Бут начал анализировать возможности механизации двуязычного словаря, а позже стал сотрудничать с Ричардом Хуком Риченсом (*Richard Hook Richens*, 1919–1984), который использовал перфокарты для дословного перевода аннотаций к научным статьям. После окончания Второй мировой войны, в 1949 г., У. Уивер издал меморандум, в котором описал перспективы машинного перевода и возможные методы его реализации: теорию информации Шеннона, статистические методы и пр.

Прорывным этапом в развитии машинного перевода стал 1954 г., когда совместно с компанией IBM и Джорджтаунским университетом в США был проведен эксперимент по переводу около 60 предложений химической тематики с русского языка на английский язык. В памяти системы содержалось всего шесть грамматических правил и 250 лексических единиц. Эти правила в научной литературе сводятся к следующим:

- к некоторой лексической единице подбирается точный эквивалент;
- происходит изменение порядка слов в некотором словосочетании, что характерно для англоязычных именных групп, в которых ядерный элемент располагается в конце;
- выбор лексической единицы для данной ситуации определяется левым контекстом;
- выбор лексической единицы для данной ситуации определяется правым контекстом;
- опущение в предложении на целевом языке лексической единицы, которая присутствовала в исходном языке;
- вставка в предложение на целевом языке лексической единицы, которая отсутствовала в исходном языке.

После проведения данного эксперимента в мире началось бурное развитие систем машинного перевода. Например, в СССР в 1956 г. начало свою работу объединение по машинному переводу,

а в Институте прикладной математики Ольга Сергеевна Кулагина (1931 – 2004 гг.) работала над созданием машинного переводчика с французского языка.

Наравне с достижениями в автоматизации переводческого процесса стоит отметить, что существующие на тот момент системы развивались достаточно медленно. Именно поэтому в 1964 г. в США общественность взволновал вопрос об отсутствии прогресса в области машинного перевода. Был создан Консультативный комитет по автоматической языковой обработке (Automatic Language Processing Advisory Committee, ALPAC), исследовавший функционал существующих машин и в 1966 г. выпустивший отчет, в котором утверждалось, что машинный перевод – медленный и неточный процесс, который стоит дороже, чем перевод, выполненный человеком. Это заявление привело к угасанию интереса к данному вопросу. В Советском Союзе наблюдалась аналогичная тенденция, поэтому машинный перевод развивался в небольших научных сообществах. Тем не менее в следующем десятилетии вышла первая рабочая система «Электротехнический автоматический перевод» (ЭТАП). Над этим проектом работали Ю.Д. Апресян (1930 – н.вр.), И.А. Мельчук (1932 – н.вр.) и т.д. На данный момент ЭТАП уже позиционируется как многоцелевой лингвистический процессор. В зарубежных странах в 1970–1990-х гг. также появлялись различные системы: Logos (языковые пары «немецкий-английский» и «английский-французский»), Metal (языковая пара «немецкий-английский») и др.

Современный этап развития систем машинного перевода начинается в 1990-х гг. В 1991 г. в Санкт-Петербурге была основана компания PROMT, в 1993 г. началось развитие проекта C-STAR, основной темой которого стал машинный перевод в сфере туризма. В 2000 г. была разработана система ALPH японской лабораторией ATR (языковые пары «японский-английский» и «китайский-английский»). Основной подход, заложенный в системе, – перевод на основе примеров. В связи с развитием сети «Интернет» многие разработчики перешли от разработки стационарных систем машинного перевода к онлайн-системам. В 2006 г. появляется система Google Translate, в 2011 г. – Яндекс.Переводчик, в 2017 г. – немецкая система DeepL.

3.2. Модели машинного перевода

Существует несколько основных подходов к решению задач машинного перевода. Первый подход – ***правиловый (rule-based approach)*** – является наиболее известным методом, в котором алгоритм перевода использует разноуровневую языковую информацию о входном тексте. Системы такого типа оснащены словарями и грамматиками, на основе которых производится формальный анализ морфологии, синтаксиса и семантики входного текста, генерирующий перевод на выбранный язык. Они получили свое развитие в 1950-х гг. после Джорджтаунского эксперимента.

Подвидом правиловых систем выступают ***трансферные системы,*** они в наибольшей степени подражают человеку-переводчику. К основным процедурам обработки текста в подобных системах относят следующие:

1. *Графематический анализ* – процесс разбора текста на уровне графем, т.е. на уровне отдельных букв и их сочетаний. Он включает в себя идентификацию и классификацию графем, определение правильности орфографии, исправление ошибок, а также коррекцию знаков пунктуации.

2. *Токенизация* – процесс разбиения текста на отдельные единицы. Токенизация может быть основана на различных правилах, таких как разделение по знакам препинания, пробелам или другим символам.

3. *Лемматизация* – процесс приведения слова к его базовой или словарной форме, называемой леммой. Лемматизация учитывает грамматические особенности слова и приводит их в нормализованную форму. Например, слово *бежит* будет приведено к лемме *бежать*. Лемматизация часто применяется для текстов на языках с богатой морфологией. Наряду с лемматизацией используется и *стемминг* – процесс приведения слова к его основе (псевдооснове) или корню (псевдокорню) путем удаления суффиксов и окончаний. Стемминг не учитывает грамматические правила. Например, слово *приходит* будет приведено к основе *приход*. Стемминг используется в ряде приложений, таких как поисковые системы, чтобы упростить поиск по ключевым словам.

4. *Семантико-синтаксический анализ* – комплекс процедур по созданию формальной структуры предложения с учетом их лексических значений и семантических ролей. Среди способов представления структуры предложения выделяют два основных: деревья зависимостей и деревья непосредственно составляющих (рис. 3.1, 3.2).

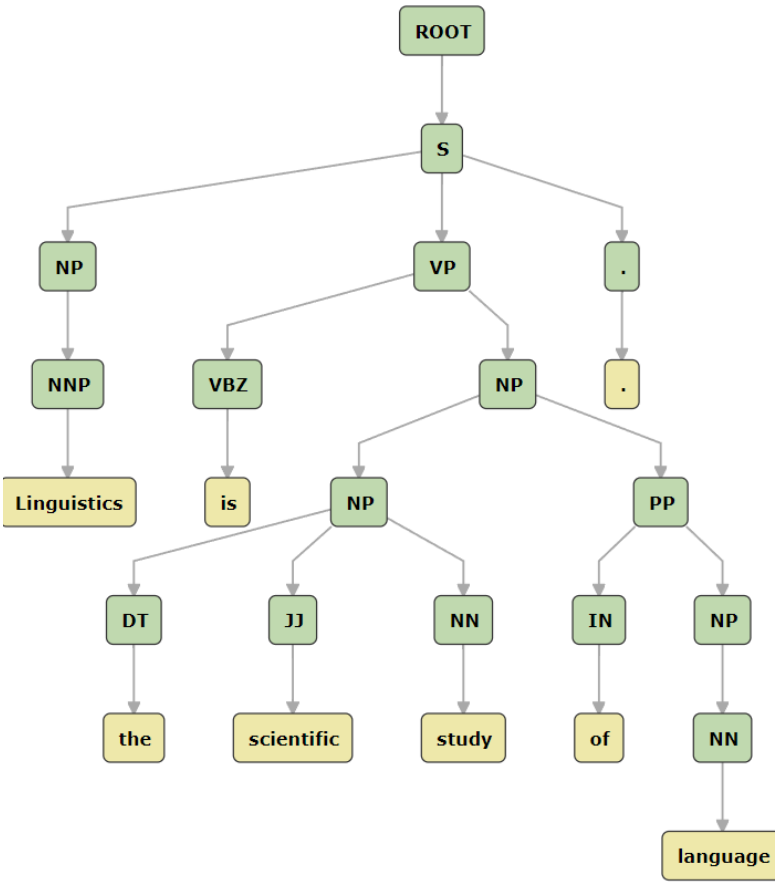

Рис. 3.1. Визуализация дерева непосредственно составляющих

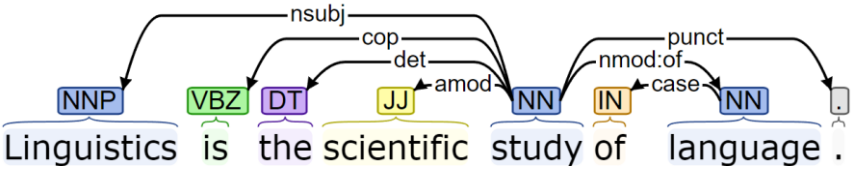

Рис. 3.2. Визуализация дерева зависимостей

Дерево зависимостей – это структура, используемая в лингвистике для представления синтаксических отношений между словами в предложении. В таком дереве каждое слово представлено узлом, а связи между ними – дугами. Связи в дереве зависимостей обычно описывают синтаксические отношения, такие как предложные отношения, атрибутивные отношения и др.

Грамматика непосредственно составляющих – это формализм, идея которого заключается в том, что предложение разделяется на части, при этом на каждом шаге деление происходит на две составляющие. В результате обычно получаются многоуровневые составляющие, которые записываются в виде скобок или в виде дерева. Результатом работы такой грамматики является ***дерево непосредственно составляющих.***

Другие подвиды правиловых систем – ***системы пословного перевода*** и ***системы на основе интерлингвы.*** Первые системы не учитывают особенности синтаксиса и семантики целевого языка. Вторые системы предназначены для тех языковых пар, для которых существует мало оцифрованной информации для создания «прямой» системы машинного перевода. Они основаны на *интерлингве* – языке-посреднике. Если в языковой паре $X - Y$ невозможно разработать качественную систему машинного перевода, то рассматривается триада $X - Z - Y$, в которой для языковых пар $X - Z$ и $Z - Y$ уже разработаны качественные системы машинного перевода.

Второй подход основан ***на примерах (example-based approach).*** В его рамках лингвистические системы охарактеризованы как системы, использующие *аналогию*. Третий подход, ***статистический (statistical approach),*** набрал популярность с конца 1980-х благодаря Исследовательскому центру Томаса Ватсона (*Thomas J. Watson Research Center*). Оба подхода используют параллельные корпусы, в которых данные о двуязычных текстах выводят некоторые закономерности, позволяющие находить соответствие лингвистических единиц для двух разных языков. Параллельные корпусы в системах, основанных на примерах, представляют собой множество примеров соответствия словосочетаний и предложений одного языка словосочетаниям и предложениям другого. При получении некоторого предложения на исходном языке оно разбивается на отдельные конструкции, соответствующие уже имеющимся в корпусе. Для этих конструкций в системе происходит поиск эквивалентов на языке перевода. Наконец, из отдельных фрагментов вновь составляется единое предложение. Для статистических систем данные о двуязычных коллекциях текстов используются для вывода лингвостатистических закономерностей (соответствием лингвистических единиц из двух разных языков).

В 2014 г. научным коллективом под руководством Дмитрия Баданау (*Dzmitry Bahdanau*) была представлена одна из первых научных работ по внедрению нейронных сетей в процесс машинного перевода, а уже с 2016 г. компания Google внедрила первую промышленную систему. ***Нейросетевой машинный перевод (neural machine translation)*** – это модель машинного перевода, основанная на использовании нейронных сетей. Системы нейросетевого перевода обучаются на больших объемах предложений и не требуют точного соответствия слов и разбивки на фразы. Вместо этого система преобразует исходное предложение в числовой набор и затем декодирует числа в слова на другом языке. При декодировании используется контекст всего исходного предложения и предыдущих предсказанных слов, что позволяет более точно выбирать слова для перевода. В процессе тренировки системы сравнивают предсказанные и правильные слова, а веса в нейросети обновляются для улучшения предсказаний. Подобный процесс применяется к каждому предложению n раз, пока система не предскажет наилучший вариант перевода. Системы нейросетевого перевода отличаются своими параметрами: в какие векторные представления кодируются слова на разных этапах, каковы размеры этих векторов, какими математическими операциями они связаны друг с другом и каким образом обеспечивается запоминание важных слов и «забывание» менее важных. Все современные системы также содержат механизм внимания, который фокусируется на словах исходного предложения при предсказании следующего слова. Механизм внимания похож на так называемый *word alignment* (поиск пословных соответствий в текстах на разных языках), но он более «мягкий» и не всегда точно показывает, какое слово было переведено. Некоторые системы, такие как *Transformer*, используют несколько уровней внимания к исходному предложению и уже предсказанному переводу (*self-attention*).

3.3. Лингвистические проблемы машинного перевода

Проблемы перевода фразеологических некомпозиционных единиц. Несмотря на все достоинства современных систем машинного перевода, такие компоненты языка, как стилистика, семантика и прагматика, в большинстве случаев остаются далеко за пределами «понимания» искусственного интеллекта. Например, при работе с идиомой *The early bird catches the worm* переводчик смело может выдать русский эквивалент: *Кто рано встает, тому бог подает*. Опытный переводчик также без проблем сможет перевести с русского языка на английский язык следующий фразеологизм: *Копейка рубль бережет – A penny saved is a penny earned*. В ситуациях, когда нет нужного эквивалента, переводчик прибегает к трансформациям. Например, известная русская пословица *Баба с возу – кобыле легче* может быть переведена так: *It will be easier without him or her*. Просторечное слово *баба* было опущено при переводе (оно несет некоторый национальный колорит), но общая идея донесена до гипотетического слушателя. Не каждый иностранец понимает пословицы и поговорки, переведенные дословно.

Рассмотрим те же самые фразы, переведенные несколькими системами машинного перевода с русского языка на английский язык и наоборот (табл. 3.1).

Таблица 3.1

Пословицы и поговорки, переведенные с помощью систем машинного перевода

Исходная пословица или поговорка	Примеры автоматического перевода		
	Google Translate	*Яндекс.Переводчик*	*PROMT.Онлайн*
The early bird catches the worm	Ранняя пташка ловит червя	Ранняя пташка червяка ловит	Ранняя пташка ловит червя
Копейка рубль бережет	Penny saves the ruble	Every penny counts	Take care of the pence, the pounds will take care of themselves
Баба с возу – кобыле легче	Baba with a cart – mare easier	A woman with a cart – the Mare is easier	The woman with a cart – a mare is easier
С миру по нитке – голому рубаха	From the world on a thread – a naked shirt	With the world on a string – bare shirt	Many a pickle makes a mickle

Английскую пословицу все три системы перевели практически одинаково, за исключением *Яндекс.Переводчика*. Вместо слова *червя* было употреблено слово *червяка*, что в принципе не меняет идеи, хотя нарушается целостность фразы. Можно сказать, что с первой фразой системы машинного перевода справились: даже при наличии русского эквивалента данное выражение также может употребляться, хотя частота использования будет гораздо ниже.

Вторая пословица была переведена уже по-разному, при этом *Яндекс.Переводчик* и *PROMT.Онлайн* предложили совершенно разные варианты, которые являются корректными, поскольку смысл передан верно. Google Translate перевел фразу дословно, хотя иностранец уже сам может логически попытаться довести ее до выражения *Every penny counts*, так как лексические единицы в обеих идиомах семантически близки.

Намного хуже обстоит дело с национальными пословицами и поговорками, так как, например, пословицу *Баба с возу – кобыле легче* все три системы перевели дословно, а в случае с пословицей *С миру по нитке – голому рубаха* аналог смогла отыскать только одна система – *PROMT.Онлайн*.

Лексико-семантические проблемы перевода. Еще один яркий пример того, что составляет проблему для переводчика, – это неологизмы и окказионализмы, которые влияют на динамику лексического состава языка. Их необходимо умело передать в другом языке. В табл. 3.2 приведены цитаты из различных произведений русской литературы. В цитатах выделены окказионализмы.

Таблица 3.2

Сравнение переведенных окказионализмов

Исходная цитата	Машинный перевод (*Google Translate*)	Человеческий перевод
«Да на что и нашему-то брату знать по-французски, на что? С барышнями в мазурке **лимонничать,** с чужими женами **апельсинничать?**» *Ф.М. Достоевский,* *«Село Степанчиково и его обитатели»*	Yes, what should our brother know in French, what for? With the ladies in mazurka **limonnic** with **orange** wives?	Is there any sense for him to learn French? Will he **flirt** with dancing ladies or **coquet** with other men's wives?
«Вокруг, с лицом, что равно годится быть и лицом и ягодицей, **задолицая** полиция». *В.В. Маяковский,* *«Владимир Ильич Ленин»*	Around, with a face that is equal to being a face and buttock, **zadolitsaya** police.	There are **ass-faced** policemen everywhere: their faces resemble normal faces and asses as well.

При переводе цитат с помощью систем машинного перевода бóльшая часть окказионализмов была передана с помощью транслитерации, так как в базе знаний системы не было информации о данных словах. Переводчик для передачи смысла слов использовал различные переводческие способы: подбор синонимов или словообразование (сложное слово из двух корней). Могут быть предложены и другие варианты перевода, но, тем не менее, данное сравнение демонстрирует преимущество человеческого перевода над машинным.

Лексические проблемы характерны не только для текстов художественной литературы, но и для узкоспециализированных текстов, в которых не соблюдается терминологическое единообразие. Так, для медицинского текста система машинного перевода Google не учла тематическую принадлежность, в результате чего для выражения *come in pairs* предложен буквальный перевод

идут парами, в то самое время как студент-выпускник (2022 г.) провел процедуру предпереводческого анализа, в результате которой он предложил верный вариант – *наследуются парами* (табл. 3.3).

Таблица 3.3

Сравнение переведенных медицинских текстов

Исходная цитата	Машинный перевод (*Google Translate*)	Человеческий перевод
Chromosomes usually come in pairs, with one chromosome from each pair coming from the father and one from the mother	Хромосомы обычно идут парами, по одной хромосоме из каждой пары от отца и одной от матери	Хромосомы обычно наследуются парами – по одной от каждого из родителей

Разрешение лексической многозначности также представляет собой одну из самых сложных задач. Эта проблема впервые была поставлена в контексте разработки систем машинного перевода во время Джорджтаунского эксперимента, для ее решения предлагалось использовать анализ левого и правого контекстов. С течением времени разрешение лексической многозначности стало ключевой задачей не только для систем машинного перевода, но и для других систем обработки естественного языка, таких как информационный поиск и классификация текстов. Существует два основных класса механизмов для разрешения многозначности в системах машинного перевода. Первый класс – автоматические механизмы, которые предлагают полностью компьютерное решение задачи. Второй класс – интерактивные (диалоговые, полуавтоматические) механизмы, которые предполагают совместное решение задачи человеком и компьютером. Подходы к автоматическому разрешению многозначности могут включать такие методы-фильтры, как правила сочетаемости лексем, правила использования ряда актантов в синтаксемах и предикативных структурах. Также могут применяться статистические методы, которые не требуют явного указания лексического значения.

Для решения проблем лексической неоднозначности в 1986 г. Майкл Леск (*Michael Lesk*, 1945 – н.вр.) предложил алгоритм, суть которого сводится к подсчету числа слов, которые встречаются как в словарном определении значения слова, так и в его контексте. Итоговым вероятным значением признается то, у которого большое количество пересечений контекста и словарного определения. Так, в англоязычном предложении *The armoured fighting vehicle hit a mine, it needs replacing the caterpillar* для лексической единицы *caterpillar* можно подобрать как минимум два значения:

1) the larva of a butterfly or moth, which has a segmented wormlike body with three pairs of true legs and several pairs of appendages similar to legs;

2) *an articulated steel band passing round the wheels of a vehicle for travel on rough ground*.

В контексте предложения и втором определении алгоритм находит совпадение по лексеме *vehicle*, значит, алгоритм будет интерпретировать слово *caterpillar* в данном предложении как часть транспортного средства, а система машинного перевода предложит вариант перевода *гусеничная лента*.

Для русского языка метод Леска не всегда применим в силу более свободного порядка слов. В качестве альтернативы предлагаются текстовые коллекции, размеченные лингвистами вручную. Такие коллекции обычно выглядят следующим образом: в первом столбце приводятся примеры лексически неоднозначных предложений, во втором указывается неоднозначная лексема со всеми морфосинтаксическими параметрами, а в третьем – словарное определение слова. Итоговая текстовая коллекция делится на две части: тренировочную выборку (обычно 80 процентов от всего объема данных) и тестовую выборку (обычно 20 процентов от всего объема данных). Тренировочная выборка подается на вход некоторому алгоритму машинного обучения, а на тестовой оценке проводится сам эксперимент по предсказыванию значений. На последнем этапе предсказанные результаты сравниваются с размеченными эталонными, выводится количественная оценка системы предсказания и делается вывод о ее жизнеспособности. Подобные подходы могут использоваться и для других языков, в том числе и для тех, в которых невозможно по каким-либо причинам собрать большой объем материала для эксперимента (*few-shot text classification*).

Грамматические проблемы машинного перевода. Передача эмфатических конструкций также может быть проблемой для машинного перевода. Для подтверждения разберем пример (табл. 3.4). Представим, что два закадычных друга решили пообщаться, однако в процессе разговора завязался спор. Один другому говорит:

– Ты ***ведь*** вчера ходил к начальнику, чтобы попросить повышения! Не ври мне!
– Это что? ***Камешки в мой огород***?

Таблица 3.4

Примеры перевода диалога

Человеческий перевод	Машинный перевод	Система
– You ***did*** go to the boss yesterday to ask for a promotion! Don't lie to me! – What do you mean? ***Was that aimed at me?***	– You ***went*** to the boss yesterday to ask for a raise! Do not lie to me! – What's this? ***Pebbles in my garden?***	*Google Translate*
– ***Did you go*** to a chief to ask for a promotion? Don't tell a lie to me. – What are you talking about? ***Are you trying to blame me in this case?***	– You ***went*** to the chief yesterday to ask increases! Don't lie to me! – It that? ***Stones in my kitchen garden?***	*PROMT. Online*

В этом диалоге был сделан упор на эмфатическое слово *ведь*, которое не имеет точного аналога в английском языке. Его надо перевести, используя лексико-грамматические трансформации. Также необходимо отметить, что во второй реплике был использован фразеологизм *бросать камень в огород*, что означает *ругать, поносить*. В человеческих переводах были соблюдены оба варианта перевода: слово *ведь* было выражено на языке перевода с помощью эмфатического *did*, а также с помощью преобразования восклицательного предложения в вопросительное. Фразеологический оборот в первом переводе был заменен на схожее выражение, которое используется в английском языке. Во втором переводе полностью перестроена лексико-грамматическая структура предложения: *to blame* – винить, *the Present Continuous Tense* – показатель раздражения.

Машинный перевод не справился с данной задачей. Во-первых, слово *повышение* в значении *продвижение по карьерной лестнице* было переведено как *raise* (оно было бы оправдано, если бы речь шла о деньгах, но такое значение здесь отсутствует) и *increase* (увеличение в размерах). Фраза *Это что?* была переведена дословно. Система машинного перевода *PROMT.Онлайн* с точки зрения грамматики не смогла правильно построить эту фразу, так как в ней отсутствует сказуемое. Более того, указанные системы перевели фразеологический оборот буквально.

Пользователи также сталкиваются с проблемами перевода при попытке передачи нестандартных атрибутивно-именных групп. Общее правило перевода таких групп в языковой паре «английский язык–русский язык» можно описать как «*12**3**–**3**12*». Это означает, что конечный элемент англоязычной атрибутивно-именной группы становится начальным в русском языке: *Black Sea* ***Coast – побережье*** *Черного моря*. Отметим, что для узкоспециализированных текстов атрибутивно-именная группа, которая является многокомпонентным термином, может быть передана иначе. В табл. 3.5 представлен перевод медицинского многокомпонентного термина: *1p* – название, поэтому в русском языке это название будет левосторонним приложением при слове «делеция». Переводчик учел эту особенность, а система машинного перевода не справилась с задачей.

Таблица 3.5

Пример перевода атрибутивно-именной группы

Исходная цитата	Машинный перевод (*Google Translate*)	Человеческий перевод
1p interstitial deletions	1p интерстициальные делеции	Интерстициальные делеции 1p

Приведенные примеры не охватывают всех возможных проблем перевода, так как для создания полной классификации ошибок необходимо провести эксперименты с привлечением данных, размеры которых будут соизмеримы с размерами данных, представленных в национальных и вебкорпусах. Отметим, что к проблемам машинного перевода также относят проблемы выравнивания текста, отсутствие переведенных в электронную форму некоторых языков мира, отсутствие письменности некоторых языках (например, бесписьменный язык чунг на северо-западе Камеруна), а также особенности артикуляции человека и фоновые шумы при работе с системами устного машинного перевода.

3.4. Постредактирование текстов машинного перевода

Ряд лингвистических ошибок, которые встречаются в текстах машинного перевода, привели к развитию отдельного направления в переводоведении – ***постредактирования машинного перевода,*** т.е. улучшения качества переведенного системой текста посредством устранения лингвистических ошибок. Говоря бытовым языком, постредактирование – это перевод с «неверного механического» языка X на «верный человеческий» язык X. Сразу необходимо отметить, что постредактирование отличается от редактирования, так как постредактирование машинного перевода подразумевает корректировку результата деятельности автоматизированной системы, а редактирование – исправление текста, предложенного не системой, а переводчиком.

Сама идея данного направления просматривалась при появлении первых работающих систем машинного перевода. Человек, владеющий целевым языком (в системе П.П. Смирнова–Троянского) и занимающийся восстановлением морфосинтаксических связей переведенных лексических единиц в тексте, являлся как раз постредактиром. Как профессия само направление было упомянуто в работах Мюриэла Васконселлоса в середине 1980-х гг.

Специалист, занимающийся корректировкой заранее переведенного системой машинного перевода текста, называется ***постредактор.*** Согласно мнению М.А. Ивлевой и В.А. Хасановой, постредакторами могут быть переводчики, которые не только в совершенстве владеют целевыми языками и языками перевода, но и осведомлены о внутренней работе систем машинного перевода. Другими словами, современному специалисту желательно не только иметь лингвистическое образование, но и техническое. Данное требование, по мнению Н.В. Нечаевой, – одно из базовых наравне с практическим опытом перевода и постредактирования текстов. Ведь в зависимости от модели машинного перевода (статистическая, основанная на правилах или нейросетевая) и жанра текста постредактор будет выбирать различные стратегии изменения текста.

Сегодня не существует полного ряда правил постредактирования текстов для всех возможных языковых пар, однако самые частотные из них описаны в руководствах и стандартах. Например, требования, описанные в международном стандарте ISO 18587:2017, почти аналогичны требованиям, предъявляемым к традиционному переводу (международный стандарт ISO 17100:2015). Приведем некоторые из них.

1. Текст перевода должен быть адекватным и соотноситься с оригинальным текстом по смыслу.
2. В тексте перевода необходимо следить за единством терминологии.
3. Текст перевода должен учитывать специфику культуры целевого языка, что приводит к адаптации некоторых компонентов текста (единицы измерения, оформление дат, валюты и пр.).
4. Текст перевода должен придерживаться норм целевого языка.
5. Стиль текста должен соответствовать его виду.

В руководстве для постредакторов, выпущенном ассоциацией TAUS (Translation Automation User Society, «Общество пользователей автоматизации перевода»), выделяются как легкое (light), так и полное (full) постредактирование, при этом каждый подвид описывает свои правила постредактирования. При легком постредактировании важно обращать внимание на верное употребление терминологии, соблюдение грамматических норм целевого языка и отсутствие искажения фактов и смысла. Специалист вносит наименьшее количество правок, в связи с чем читатель в дальнейшем при знакомстве с таким текстом сможет уловить лишь основную тематическую составляющую текста. При полном постредактировании ряд правил расширяется. Помимо указанных требований, необходимо совершить предпереводческий анализ, провести локализацию определенных компонентов текста, внести стилистические корректировки и пр.

Наконец, количество внесенных изменений в текст напрямую зависит от качества работы самой системы. На сегодняшний день специалисты-переводчики разработали разнообразные критерии определения качества машинного перевода:

- подсчет процента совпадающих слов в отредактированном и неотредактированном переводе;
- создание шкалы для оценки адекватности, понятности перевода, а также соответствия смысла оригиналу;
- получение от читателя ответов на вопросы по тексту, учет эквивалентности на разных языковых уровнях.

Все существующие подходы к оценке качества работы систем машинного перевода можно разделить в зависимости от области исследования. В компьютерной лингвистике разрабатываемые

системы оцениваются по таким параметрам, как точность (precision), полнота (recall) и F-мера (F-measure) – среднегармоническое первых двух величин. Для оценки работы машинного перевода разрабатывают и отдельные метрики. Одна из самых популярных метрик оценки качества работы систем машинного перевода – BLEU (Bilingual Evaluation Understudy) – «измерение различий между автоматическим переводом и одним или несколькими эталонными пользовательскими переводами одного исходного предложения». Основной подход – сопоставление текста машинного перевода и профессионального перевода: чем больше сходств машинного перевода с человеческим, тем выше качество.

Метрика METEOR использует так называемую «систему штрафов», начисляемых за несовпадающие фрагменты. Фрагмент – часть предложения, переводы которой совпадают у человека и программы. Для учета более длинных совпадений METEOR рассчитывает штраф, группируя слова (униграммы) в наименьшее возможное количество фрагментов. Чем длиннее *n*-граммы, тем меньше фрагментов.

Другая формула известна как Word Error Rate (WER), идея которой основана на расстоянии Левенштейна (расстоянии редактирования). Формула оценивает количество вставленных, удаленных или замененных слов в тексте:

$$WER = \frac{S + D + I}{S + D + C},$$

где S – количество замененных слов; D – количество удаленных слов; I – количество вставленных слов; C – количество верных слов. Общее количество слов в тексте $N = S + D + C$. Чем ближе значение WER к нулю, тем выше качество переведенного текста.

Отечественные лингвисты также занимаются разработкой формул для оценки качества работы систем машинного перевода. Например, С.О. Шереметьева предложила следующую формулу:

$$E = \frac{N_w}{N_{sem\,\&lex} \times 10 + N_{synt} \times 5 + N_{morp\,h} \times 3 + N_{cov} \times 2 + N_{style}},$$

где N_w – общее количество слов в тексте; $N_{sem\,\&lex}$ – количество лексико-семантических ошибок в тексте; N_{synt} – количество синтаксических ошибок в тексте; $N_{morp\,h}$ – количество морфологических ошибок в тексте; N_{cov} – количество непереведенных слов; N_{style} – количество стилистических ошибок в тексте. Чем ниже показатель E, тем выше качество машинного перевода.

Были разработаны и узконаправленные метрики для совершенно различных языков. Например, для перевода предложения с английского языка на японский язык нужно существенно изменить порядок слов, поэтому в 2010 г. была разработана метрика Rank-based Intuitive Bilingual Evaluation Score (RIBES):

$$RIBES = \frac{NSR \times p^{\alpha} + NKT \times p^{\alpha}}{2}.$$

В формуле используются коэффициенты ранговой корреляции *n*-граммов эталонного перевода и машинного перевода: NSR – нормализованные коэффициенты Спирмена; NKT – нормализованные коэффициенты Кендалла. Эти коэффициенты выступают как показатель веса для точности – p; α – параметр в диапазоне от 0 до 1. $RIBES$ – результат усреднения взвешенных точностей.

4. CAT-СИСТЕМЫ В ПЕРЕВОДОВЕДЕНИИ

CAT-система (Computer-Assisted Translation, Computer-Aided Translation) – это специальное программное обеспечение, разработанное для автоматизации процесса перевода. В отличие от систем машинного перевода, CAT-инструменты не предназначены для замены переводчика, а служат для упрощения его работы. Эти программы являются своего рода текстовыми редакторами с дополнительными функциями, которые позволяют:

- получать доступ к словарям и глоссариям прямо из рабочей среды;
- запоминать переводы для конкретных слов, словосочетаний или даже целых предложений и автоматически предлагать их использование, когда встречаются похожие фрагменты (программа также покажет процент совпадения);
- упрощать процесс форматирования текста.

Процедуру работы в таких системах можно описать следующим образом. Менеджер проекта загружает сверстанный документ в систему, настраивает его параметры, добавляет дополнительные материалы и устанавливает сроки выполнения. Затем он направляет задание исполнителям. Внутри системы документ может быть разделен между несколькими переводчиками, фрагменты, не требующие перевода, могут быть скрыты, а к процессу сразу можно привлечь редакторов или корректоров. Основная задача менеджера – проверить и исправить форматирование: система сохраняет оригинальное форматирование файла, но длина переведенного текста может отличаться от оригинала. Использование CAT-инструментов помогает сэкономить время и деньги. Переводчику не нужно искать ранее сохраненные термины в базе данных, так как они уже утверждены клиентом. Части документа, совпадающие с сегментами из памяти переводов (*translation memory*), могут быть переведены автоматически (при 100%-ном совпадении перевод можно подтвердить сразу). Это позволяет сохранить единый стиль и терминологию в текстах одного клиента.

4.1. Обзор существующих систем

SDL Language Cloud (RWS Language Cloud) – облачная среда, которая централизует процесс перевода. Этот инструмент объединяет всех участников переводческого процесса: создателей контента, менеджеров проектов, переводчиков и редакторов-корректоров. Это не полностью самостоятельное решение для переводчиков. *SDL* не предоставляет возможность использовать *Language Cloud* для перевода напрямую из браузера, но предлагает выбор: интегрировать его с *SDL Trados Studio* или использовать плагин для *Microsoft Word*. За плату пользователь получает следующие возможности:

- интеграция *Language Cloud* с *SDL Trados Studio*;
- просмотр баз терминов;
- поиск похожих слов и выражений (независимо от словоформы);
- лингвистический поиск (на основе схожести корневых морфем);
- управление терминами в базе данных терминов;
- импорт/экспорт терминосистем в форматах Microsoft Excel, CSV (Comma-Separated Values) и др.;
- предоставление доступа к терминосистеме другим пользователям в рамках вашего аккаунта *Language Cloud*.

Эта подписка не включает машинный перевод и является базовым планом для частных пользователей. Цены на *Language Cloud* для бизнеса *SDL* предоставляет только по запросу. Корпоративная подписка отличается возможностью создания собственных определений для терминов в базе и предоставления доступа к терминосистеме через ссылку без подключения к вашему аккаунту. Для корпоративных клиентов существует версия *Trados Studio Professional Network*, в которой лицензии находятся на сервере *SDL*. Стоимость такой версии зависит от количества лицензий. *Trados Studio* часто указывают в переводческих вакансиях, но работа с этим инструментом может оказаться затруднительной для начинающего переводчика из-за обилия функций.

SDL Language Cloud и *Trados Studio Professional Network* лучше всего подходят для удаленных команд по переводу и локализации с большим объемом заданий на перевод, которые уже используют услуги *SDL* и предпочитают оставить управление сетевой стороной *SDL*.

Другой инструмент – ***memoQ.*** Система существует как отдельная программа с 2006 г. Компания *Kilgray* предлагает пользователям сразу несколько продуктов (рис. 4.1).

Самым ранним решением компании стала ***memoQ translator pro*** – система для упрощения переводческого процесса. Программа обладает широким набором функций, которые на первый взгляд не уступают *SDL Trados*. Компания предлагает *iOS*-приложение для перевода более чем на 30 языков (с режимом диктовки текста), модуль для управления проектами, облачные и серверные решения *memoQ*.

Инструмент ***memoQ cloud*** не требует дополнительной установки программного обеспечения, облачная версия может использоваться как вместе с установленным *translator pro*, так и напрямую из браузера. Облачная версия не предназначена для частного пользования. Для нее необходимо иметь собственный сервер, который следует указать при первом входе. Инструмент ***memoQ TMS*** – это система управления для *memoQ cloud*, суть которой состоит в интеграции продуктов и решений *memoQ* в инфраструктуру компаний.

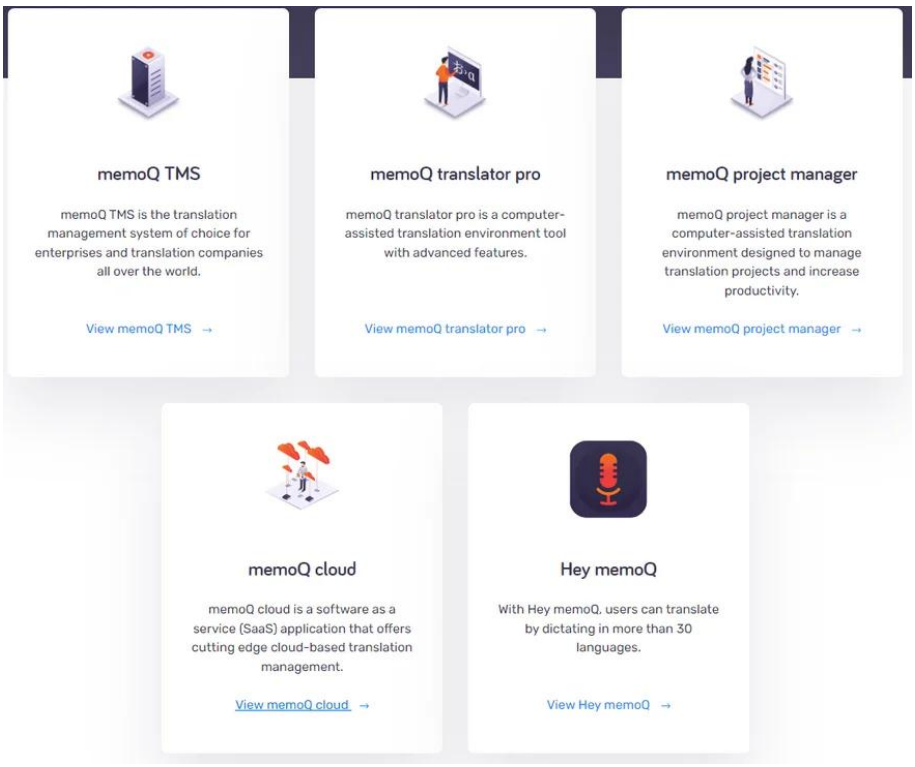

Рис. 4.1. Семейство продуктов *memoQ*

Основные возможности *memoQ cloud*:

- оплата в зависимости от количества пользователей;
- взаимодействие членов переводческой команды в реальном времени на платформе;
- безопасность и резервное копирование данных;
- гибкое управление лицензиями: можно создавать лицензии для *translator pro*, для работы в браузере или их комбинацию;
- работа с модулем *QTerm*, который регулирует терминологию проекта;
- интеграция с переводческим порталом *Language Terminal*.

Инструмент *memoQ* может использоваться компаниями, которым приходится сталкиваться с разными форматами данных при выполнении перевода или которые хотят создать целостную переводческую инфраструктуру как на основе сторонних серверов, так и на базе собственных.

Знакомство с переводческой работой часто начинается с инструмента ***Smartcat.*** Он был создан в 2012–2015 гг. как внутренний переводческий инструмент компании ABBYY. Некоторое время *Smartcat* был связан с ABBYY, но вскоре стал независимым и начал стремительно развиваться. Сейчас *Smartcat* представляет собой не только систему автоматизации перевода, но и маркетплейс для фрилансеров, заказчиков перевода и специализированных бюро. *Smartcat* также имеет свою систему интеграции с различными современными инструментами, включая машинный перевод, подсказки при переводе и систему расчетов между заказчиками и исполнителями. *Smartcat* полностью бесплатен для всех пользователей, но также имеются платные услуги для расширения возможностей. Существуют различные пакеты услуг: Forever Free, Starter, Unite, Enterprise. Основные функции *Smartcat*:

- неограниченное количество пользователей, памяти перевода и баз терминов;
- доступ к более чем 10 сервисам машинного перевода;
- интеграция с внешними сервисами, такими как Google Drive;
- наличие среды, которая обеспечивает взаимодействие всех участников переводческого процесса;
- управление исполнителями;
- использование API (Application Programming Interface);
- выплаты денежных средств исполнителям;
- доступ к маркетплейсу (площадке для размещения вакансий и поиска исполнителей, *Smartcat* удерживает 10 процентов от заказа исполнителя в качестве комиссии);
- составление отчетности (включая налоговую).

Данный функционал *Smartcat* подходит для компаний, предоставляющих переводческие услуги. Его основной недостаток состоит в том, что он является онлайн-инструментом и не имеет офлайн-версии. Это означает, что все исполнители должны иметь постоянное Интернет-соединение с достаточной пропускной способностью для работы над переводом. На сегодняшний день *Smartcat* не адаптирован для работы с мобильного устройства.

Smartcat подойдет переводческим компаниям, которые хотят напрямую находить исполнителей для проекта и вести с ними переговоры.

Описанные решения не являются единственными в сфере автоматизации переводческого процесса. Компания *STAR* разработала продукт ***Transit NXT*** – специальный инструмент для локализации приложений и сайтов. Одна из ключевых особенностей программы – возможность сохранения перевода не отдельного выражения, а всего контекста, что облегчает работу с маркетинговыми текстами, при переводе которых первостепенным параметром становится стилистика. ***Matecat*** – онлайн-среда для перевода, ее отличительные особенности – возможность выполнения перевода без регистрации, а также использование общедоступной памяти перевода (*Public Translation Memory*), которая сохраняет все переведенные когда-либо сегменты текстов. Общедоступная память переводов может стать подспорьем при работе со сложными фрагментами текста, но необходимо помнить, что в зависимости от различных лингвистических факторов такую «подсказку» нужно будет адаптировать под общий стиль итогового перевода.

4.2. Общие рекомендации по составлению глоссариев

Составление и пополнение глоссариев с использованием специализированных терминов – распространенная практика для переводчиков. Глоссарий включает термины и их переводы на один или несколько языков, а также комментарии и примеры использования. Глоссарии условно можно разделить на несколько типов:

- ***отраслевой глоссарий,*** в который добавляются термины, характерные для конкретной отрасли (например, ракетостроение);
- ***клиентский глоссарий,*** содержащий термины, характерные для организации-работодателя; такой глоссарий особенно полезен при долгосрочном сотрудничестве с фирмой;
- ***глоссарий проекта,*** который помогает поддерживать единообразие терминологии в рамках отдельного проекта.

Независимо от типа разрабатываемого глоссария, переводчику необходимо придерживаться следующих правил. Во-первых, *термины не должны дублироваться*. Повторы терминов могут ввести переводчика в заблуждение, поэтому необходимо избегать дублетов. Во-вторых, *глоссарий не должен быть избыточным*. Высокая частотность употребления лексической единицы не является основным фактором для включения слов в глоссарий. Рекомендуется добавлять только термины, характерные для продукта или компании. В-третьих, для лучшего усвоения термина предметной области *нужно указывать контекст*. В-четвертых, глоссарий *может содержать список «терминов, которые остаются без перевода» (not to be translated terms, NTBT)*. Заказчик может потребовать от переводчика не передавать на целевой язык ряд терминов (например, названия компаний). Наконец, *необходимо снабжать термины определениями*, которые помогут переводчику лучше ориентироваться в тексте.

В CAT-системах глоссарии обычно составляются и сохраняются в форматах, совместимых с самой системой CAT. Приведем некоторые из этих форматов.

TBX (TermBase eXchange) – это стандартный формат для обмена терминологическими глоссариями и базами данных. Он поддерживается большинством CAT-систем, в том числе *SDL Trados Studio* и *memoQ*.

CSV (Comma-Separated Values) – простой текстовый формат с запятой в качестве разделителя. Он широко поддерживается большинством CAT-систем и может быть легко импортирован и экспортирован.

XLSX (Microsoft Excel) – формат электронных таблиц, обрабатываемый средой Microsoft Excel. Он может быть использован для создания и редактирования глоссариев на стационарном компьютере перед их импортом в CAT-системы.

MultiTerm – формат, используемый в CAT-системе *SDL Trados Studio* для сохранения терминов и их переводов. Он имеет расширение *.sdltb*, его можно как импортировать в *Trados Studio*, так и экспортировать из него.

XML (eXtensible Markup Language) – универсальный формат для обмена данных в структурированной форме. XML обеспечивает общий формат для передачи и интерпретации данных различного рода.

В каждой CAT-системе могут быть специфические форматы или расширения файлов, поэтому лучше ознакомиться с документацией и руководствами по конкретной CAT-системе для получения дополнительной информации о форматах глоссариев.

ПРАКТИЧЕСКИЕ ЗАДАНИЯ

К разделу 1

Задание 1.1. Зайдите на сайт Voyant. Каждый студент заранее должен найти небольшой англоязычный текст приблизительно на 6000 символов по технической (физика, химия…) или гуманитарной (лингвистика, литературоведение…) тематике. Скопируйте и вставьте текст в заглавное окно, которое появится на сайте. В верхней панели найдите вкладку *Terms*. Выберите из нее слова, которые могут, на ваш взгляд, выступать терминами выбранной вами области, с абсолютной частотой не менее трех. Предложите свой перевод этих терминов на русский язык (можете свериться с необходимыми словарями). В нижней вкладке *Phrases* выберите двух- или трехсложные сочетания, которые также могут быть терминами вашей области. Предложите их русскоязычный аналог.

Задание 1.2. Зайдите на сайт Stanza. Вставьте ваш текст в заглавное окно, в графе *Annotations* выберите *Named Entities*. Посмотрите на результаты анализа: есть ли в вашем тексте прецизионная лексика? Если есть, выпишите ее и предложите русскоязычный перевод.

Задание 1.3. Выберите любой текст по технической или гуманитарной тематике (около 6000 знаков) на английском языке. Проведите стандартный предпереводческий анализ. Выпишите из текста все термины. Зайдите на сайт Национального корпуса русского языка. Введите термины (поочередно) в поисковую строку. Проанализируйте результаты. Выпишите все предложенные варианты перевода (даже если их несколько). Укажите, какой из вариантов чаще встречается в корпусе (в абсолютной частоте). Если нет никаких выдач, то предложите термин для перевода, укажите источник, из которого Вы его взяли.

К разделу 2

Задание

1. Определите жанр аудиовизуального материала.
2. Выявите систему языковых средств исходного языка.
3. Составьте глоссарий, соответствующий жанру аудиовизуального произведения.
4. Выполните перевод видеофрагмента, используя программу для субтитрирования.
5. Результат сохраните в формате *SubRip (*srt).*

Монологи общей тематики

2.1 **2.2** **2.3** **2.4** **2.5** **2.6**

2.7 **2.8** **2.9** **2.10**

Диалоги общей тематики

2.11  2.12  2.13  2.14  2.15 

Научно-технический аудиовизуальный перевод

2.16  2.17  2.18  2.19  2.20 

К разделу 3

Задание 3.1. С помощью системы PROMT.ONE переведите на русский язык данный преподавателем текст (приблизительный объем – 1800 знаков). Результаты работы оформите в виде таблицы (предложение на языке оригинала, предложение на языке перевода), в которой необходимо разметить проблематичные русско-английские сегменты переводов разным цветом в зависимости от типа ошибки следующим образом:

- лексико-семантическую ошибку (неправильный выбор значения лексемы внутри одной части речи, нарушение лексической сочетаемости слов и т.д.) – красным;
- синтаксическую ошибку (неправильные предлоги, порядок слов, залог и т.д.) – голубым;
- морфологическую ошибку (неправильный выбор морфологической формы лексемы, например при согласовании, и т.д.) – желтым;
- стилистическую ошибку (речевая избыточность, наличие речевых штампов и т.д.) – зеленым;
- непереведенные слова – серым.

Составьте диаграмму распределения ошибок в тексте. Рассчитайте количественные показатели качества машинного перевода по формуле WER и по формуле, предложенной С.О. Шереметьевой. Кратко прокомментируйте результаты работы системы машинного перевода. Составьте отчет по заданию.

Задание 3.2. Сделайте полное качественное постредактирование «сырого» текста, полученного при выполнении задания 3.1. Расширьте вашу таблицу еще на один столбец, в котором вы будете приводить исправленный вариант перевода. Выделите отредактированные фрагменты текста тем же цветом, каким были выделены неверно переведенные фрагменты во втором столбце. Составьте отчет по заданию.

Задание 3.3. Сделайте полное качественное постредактирование текста, полученного при автоматическом переводе на русский язык.

Оригинальный текст

Here's why St Petersburg is the perfect gateway to travel around Russia

Often considered Russia’s "cultural capital," St Petersburg is undoubtedly a dazzling city. It would be worth visiting for the palaces alone – or in fact, the grand museums, buzzing art scene or spectacular canals.

But many travellers are also choosing "Piter", as it’s nicknamed, as a jumping-off point for exploring the rest of Russia. It’s one of the key cities included in the north-western Silver Necklace route, and as few tourists want to miss it, is also often tacked on to classic Golden Ring tours. Both routes are diverse enough to include destinations suitable for every season, so whether it’s wintry cityscapes or sun-soaked scenery you have in mind, you’ll find them on these circuits.

To really get a feel for the essence of St Petersburg, begin with a visit to the magnificent Palace Square. Here you'll be able to tour the historic Winter Palace, with its striking green, gold and white exterior and fantastically ornate halls. The enormous State Hermitage Museum is also based here, housing one of the greatest art collections in the world, so set aside some time to admire the Picassos and Rembrandts, Michelangelo's Crouching Boy sculpture, the Rothschild Fabergé egg-shaped clock and much more. Another absolutely-must-see in St Petersburg is the Church of the Savior on the Spilled Blood, an incredibly lavish, five-domed church in the classic Russian Orthodox style. Tsar Alexander II was assassinated here in 1881, hence the evocative name.

For evening entertainment, an opera or ballet at the Mariinsky Theatre, where almost all of Russia's greatest dancers and musicians have performed over the years, makes for a memorable outing. But nightlife-wise, St Petersburg really comes into its own during the summer "White Nights", when the city has round-the-clock daylight, and there are outdoor festivals, concerts and parties held through the small hours.

Машинный перевод, предложенный системой Google Translate

Вот почему Санкт-Петербург – идеальные ворота для путешествий по России

Санкт-Петербург, который часто называют «культурной столицей» России, несомненно, великолепный город. Стоит посетить только дворцы - или даже грандиозные музеи, шумные художественные сцены или впечатляющие каналы.

Но многие путешественники также выбирают "Питер", как его называют, как отправную точку для знакомства с остальной Россией. Это один из ключевых городов, включенных в северо-западный маршрут Серебряного ожерелья, и, поскольку немногие туристы хотят его пропустить, также часто присоединяют к классическим турам по Золотому кольцу. Оба маршрута достаточно разнообразны, чтобы включать пункты назначения, подходящие для любого времени года, поэтому, будь то зимние городские пейзажи или залитые солнцем пейзажи, вы найдете их на этих трассах.

Чтобы по-настоящему почувствовать суть Санкт-Петербурга, начните с посещения великолепной Дворцовой площади. Здесь вы сможете совершить поездку по историческому Зимнему дворцу с его ярким зеленым, золотым и белым фасадом и фантастически богато украшенными залами. Здесь также находится огромный Государственный Эрмитаж, в котором хранится одна из величайших коллекций произведений искусства в мире, поэтому выделите время, чтобы полюбоваться картинами Пикассо и Рембрандта, скульптурой «Крадущийся мальчик» Микеланджело, часами в форме яйца Ротшильда Фаберже и многим другим. Еще одно место, которое обязательно нужно посетить в Санкт-Петербурге, – это Храм Спаса-на-Крови, невероятно роскошный пятиглавый храм в классическом русском православном стиле. В 1881 году здесь был убит царь Александр II, отсюда и название, вызывающее воспоминания.

Как вечернее развлечение вы можете посетить оперу или балет в Мариинском театре, где на протяжении многих лет выступали почти все величайшие танцоры и музыканты России, и это станет незабываемым событием. Но с точки зрения ночной жизни Санкт-Петербург действительно проявляет себя в летние «белые ночи», когда в городе круглосуточно светит день, а с утра проходят фестивали, концерты и вечеринки на открытом воздухе.

К разделу 4

Задание. Выполните перевод предложенных текстов в некоторой CAT-системе. Распределите роли в группе: назначьте заказчика, переводчиков отдельных сегментов и редактора. Разработайте электронный глоссарий для выбранной CAT-системы, загрузите его для унификации терминологии переводчиками. При работе с текстом обращайте внимание на теги, отвечающие за форматирование переведенного текста.

4.1. Tips for Learning a New Translation Tool

by Milica Dragićević

Do you use translation tools?
No, I don't.

This was the answer from more than 50% of respondents to my poll on the use of translation tools.

The results of the poll were based on the responses of more than 50 sworn translators and interpreters from Serbia. Although the sample size is not large enough to make generalizations, the responses do beg the question: why are translators reluctant to apply translation technologies in their translation practice?

More than 60% of my respondents said that they have difficulty learning to use them. The steep learning curve is simply not worth the hassle.

I harbored the same worries during my studies, so I can relate to my fellow translators. But one point needs to be stressed. CAT tools are developed with linguists in mind. You do not need to be a Silicon Valley-type of tech geek to be able to master them.

If you have had the same thoughts as my colleagues, here are some tips for learning a new translation tool to help you on your road to a more tech-savvy translation career.

1: Get formal training

Most translation tool providers have training programs. The formal training they offer is one of the most efficient ways to learn a new translation tool.

These training programs, like the Memsource Certification Program, contain comprehensive and systematic know-how and guide you through the step-by-step process of mastering a specific tool.

Formal training by translation tool providers is adapted to your level of knowledge. Basic level courses give you an overview of options and functions while the advanced level courses help you become proficient in using more complex functionalities.

2: Leverage the resources on offer

If for any reason you cannot attend formal training, you can learn a new translation tool by watching tutorials, videos, and webinars. You can find these resources on the translation tool provider's website, help center, and even their social media pages.

You should also go through help documentation and tips and tricks articles. They contain more detailed information on the various functions of the tool.

Last but not least, community forums are a great place to ask questions, learn more about a specific tool and connect with the community of users all around the world.

3: Organise your learning process

Translation tools can have complex functionalities developed for advanced level users, not only for linguists but also for project managers. I admit that they can be slightly off-putting to beginners. However, don't overwhelm yourself by trying to learn all there is to know right away. Simplify the learning process by mastering the basics first.

Here are some fundamentals every beginner should be familiar with:

- A basic understanding of the interface of the tool.
- How to create translation projects (importing translation files, translation memories, and term bases).
- How to open translation projects, deliver them, and how to export translation files to their target format.
- The basic user actions in the translation editor (confirming a segment, searching a translation memory and term base).

4: Attend workshops in your town

The internet, with its limitless resources, gives us the opportunity to learn almost anything we want. Although e-learning has its advantages, it does have one disadvantage. Your computer is your only learning partner. You are on your own.

At traditional workshops, however, you are surrounded by your colleagues who are facing the same challenges as you. You can help and support each other. The face-to-face interaction with your teacher and peers creates a unique sense of community you cannot develop in any other way.

I recently attended a Memsource workshop in Belgrade organized by Prima prevodi translation agency.

I had an amazing time, learned useful tips and tricks, and I met wonderful colleagues. I recommend attending translation tool workshops if you get the chance.

I hope these tips and tricks encourage you to take on the challenge of mastering a translation tool and start using it in your daily work.

4.2. History of Television

Few inventions have had as much effect on contemporary American society as television. Before 1947 the number of U.S. homes with television sets could be measured in the thousands. By the late 1990s, 98 percent of U.S. homes had at least one television set, and those sets were on for an average of more than seven hours a day. The typical American spends (depending on the survey and the time of year) from two-and-a-half to almost five hours a day watching television. It is significant not only that

this time is being spent with television but that it is not being spent engaging in other activities, such as reading or going out or socializing.

EXPERIMENTS

Electronic television was first successfully demonstrated in San Francisco on Sept. 7, 1927. The system was designed by Philo Taylor Farnsworth, a 21-year-old inventor who had lived in a house without electricity until he was 14. While still in high school, Farnsworth had begun to conceive of a system that could capture moving images in a form that could be coded onto radio waves and then transformed back into a picture on a screen. Boris Rosing in Russia had conducted some crude experiments in transmitting images 16 years before Farnsworth's first success. Also, a mechanical television system, which scanned images using a rotating disk with holes arranged in a spiral pattern, had been demonstrated by John Logie Baird in England and Charles Francis Jenkins in the United States earlier in the 1920s. However, Farnsworth's invention, which scanned images with a beam of electrons, is the direct ancestor of modern television. The first image he transmitted on it was a simple line. Soon he aimed his primitive camera at a dollar sign because an investor had asked, "When are we going to see some dollars in this thing, Farnsworth?"

EARLY DEVELOPMENT

RCA, the company that dominated the radio business in the United States with its two NBC networks, invested $50 million in the development of electronic television. To direct the effort, the company's president, David Sarnoff, hired the Russian-born scientist Vladimir Kosma Zworykin, who had participated in Rosing's experiments. In 1939, RCA televised the opening of the New York World's Fair, including a speech by President Franklin Delano Roosevelt, who was the first president to appear on television. Later that year RCA paid for a license to use Farnsworth's television patents. RCA began selling television sets with 5 by 12 in (12.7 by 25.4 cm) picture tubes. The company also began broadcasting regular programs, including scenes captured by a mobile unit and, on May 17, 1939, the first televised baseball game between Princeton and Columbia universities. By 1941 the Columbia Broadcasting System (CBS), RCA's main competition in radio, was broadcasting two 15-minute newscasts a day to a tiny audience on its New York television station.

Early television was quite primitive. All the action at that first televised baseball game had to be captured by a single camera, and the limitations of early cameras forced actors in dramas to work under impossibly hot lights, wearing black lipstick and green makeup (the cameras had trouble with the color white). The early newscasts on CBS were "chalk talks," with a newsman moving a pointer across a map of Europe, then consumed by war. The poor quality of the picture made it difficult to make out the newsman, let alone the map. World War II slowed the development of television, as companies like RCA turned their attention to military production. Television's progress was further slowed by a struggle over wavelength allocations with the new FM radio and a battle over government regulation. The Federal Communications Commission's (FCC) 1941 ruling that the National Broadcasting Company (NBC) had to sell one of its two radio networks was upheld by the Supreme Court in 1943. The second network became the new American Broadcasting Company (ABC), which would enter television early in the next decade. Six experimental television stations remained on the air during the war one each in Chicago, Philadelphia, Los Angeles, and Schenectady, N.Y., and two in New York City. But full-scale commercial television broadcasting did not begin in the United States until 1947.

THE BEGINNING OF COMMERCIAL TELEVISION

By 1949 Americans who lived within range of the growing number of television stations in the country could watch, for example, *The Texaco Star Theater* (1948), starring Milton Berle, or the children's program, *Howdy Doody* (1947Ð60). They could also choose between two 15-minute newscasts *CBS TV News* (1948) with Douglas Edwards and NBC's *Camel News Caravan* (1948) with John Cameron Swayze (who was required by the tobacco company sponsor to have a burning cigarette always visible when he was on camera). Many early programs such *as Amos 'n' Andy* (1951) or *The Jack Benny Show* (1950-65) were borrowed from early television's older, more established Big Brother: network radio. Most of the formats of the new programs newscasts, situation comedies, variety shows, and dramas were borrowed from radio, too (see radio broadcasting and television programming). NBC and CBS took

the funds needed to establish this new medium from their radio profits. However, television networks soon would be making substantial profits of their own, and network radio would all but disappear, except as a carrier of hourly newscasts. Ideas on what to do with the element television added to radio, the visuals, sometimes seemed in short supply. On news programs, in particular, the temptation was to fill the screen with "talking heads," newscasters simply reading the news, as they might have for radio. For shots of news events, the networks relied initially on the newsreel companies, whose work had been shown previously in movie studios. The number of television sets in use rose from 6,000 in 1946 to some 12 million by 1951. No new invention entered American homes faster than black and white television sets; by 1955 half of all U.S. homes had one.

McCARTHYISM

In 1947, the House Committee on Un-American Activities began an investigation of the film industry, and Sen. Joseph R. McCarthy soon began to inveigh against what he claimed was Communist infiltration of the government. Broadcasting, too, felt the impact of this growing national witch-hunt. Three former members of the Federal Bureau of Investigation (FBI) published "Counterattack: The Newsletter of Facts on Communism," and in 1950 a pamphlet, "Red Channels," listed the supposedly Communist associations of 151 performing artists. Anti-Communist vigilantes applied pressure to advertisers the source of network profits. Political beliefs suddenly became grounds for getting fired. Most of the producers, writers, and actors who were accused of having had left-wing leanings found themselves blacklisted, unable to get work. CBS even instituted a loyalty oath for its employees. Among the few individuals in television well positioned enough and brave enough to take a stand against McCarthyism was the distinguished former radio reporter Edward R. Murrow. In partnership with the news producer Fred Friendly, Murrow began *See It Now*, a television documentary series, in 1950. On Mar. 9, 1954, Murrow narrated a report on McCarthy, exposing the senator's shoddy tactics. Of McCarthy, Murrow observed, "His mistake has been to confuse dissent with disloyalty." A nervous CBS refused to promote Murrow and Friendly's program. Offered free time by CBS, McCarthy replied on April 6, calling Murrow "the leader and the cleverest of the jackal pack which is always found at the throat of anyone who dares to expose Communist traitors." In this TV appearance, McCarthy proved to be his own worst enemy, and it became apparent that Murrow had helped to break McCarthy's reign of fear. In 1954 the U.S. Senate censured McCarthy, and CBS's "security" office was closed down.

THE GOLDEN AGE

Between 1953 and 1955, television programming began to take some steps away from radio formats. NBC television president Sylvester Weaver devised the "spectacular," a notable example of which was *Peter Pan* (1955), starring Mary Martin, which attracted 60 million viewers. Weaver also developed the magazine-format programs *Today*, which made its debut in 1952 with Dave Garroway as host (until 1961), and *The Tonight Show*, which began in 1953 hosted by Steve Allen (until 1957). The third network, ABC, turned its first profit with youth-oriented shows such as *Disneyland*, which debuted in 1954 (and has since been broadcast under different names), and *The Mickey Mouse Club*.

The programming that dominated the two major networks in the mid-1950s borrowed heavily from another medium: theater. NBC and CBS presented such noteworthy, and critically acclaimed, dramatic anthologies as *Kraft Television Theater* (1947), *Studio One* (1948), *Playhouse 90* (1956), and *The U.S. Steel Hour* (1953).

TELEVISION AND POLITICS

Television news first covered the presidential nominating conventions of the two major parties, events then still at the heart of America politics, in 1952. The term "anchorman" was used, probably for the first time, to describe Walter Cronkite's central role in CBS's convention coverage that year. In succeeding decades these conventions would become so concerned with looking good on television that they would lose their spontaneity and eventually their news value. The power of television news increased with the arrival of the popular newscast, *The Huntley-Brinkley Report*, on NBC in 1956 (see Huntley, Chet, and Brinkley, David). The networks had begun producing their own news film. Increasingly, they began to compete with newspapers as the country's primary source of news (see journalism).

THE THREE NETWORKS AT THE HEIGHT OF THEIR POWER

In 1964, color broadcasting began on prime-time television. The FCC initially approved a CBS color system, then swung in RCA's favor after Sarnoff swamped the marketplace with black-and-white sets compatible with RCA color (the CBS color system was not compatible with black-and-white sets and would have required the purchase of new sets). During the 1960s and 1970s a country increasingly fascinated with television was limited to watching almost exclusively what appeared on the three major networks: CBS, NBC, and ABC. These networks purchased time to broadcast their programs from about 200 affiliates stations in each of the major cities or metropolitan areas of the United States. In the larger cities, there might also be a few independent stations (mostly playing reruns of old network shows) and perhaps a fledgling public broadcasting channel. Programming on each of the three networks was designed to grab a mass audience. Network shows therefore catered, as critics put it, to the lowest common denominator. James Aubrey, president of CBS television, doubled the network's profits between 1960 and 1966 by broadcasting simple comedies like *The Beverly Hillbillies* (1962–1971). In 1961, Newton Minow, then chairman of the FCC, called television a "vast wasteland." Programming became a little more adventurous with the arrival of more realistic situation comedies, beginning with CBS's *All in the Family* in 1971 (broadcast until 1979). Along with situation comedies usually a half-hour focused on either a family and their neighbors or a group of co-workers the other main staple of network prime-time programming has been the one-hour drama, featuring the adventures of police, detectives, doctors, lawyers, or, in the early decades of television, cowboys. Daytime television programming consisted primarily of soap operas and quiz shows until the 1980s, when talk shows discussing subjects that were formerly taboo, such as sexuality, became popular.

The three major networks have always been in a continual race for ratings and advertising dollars. CBS and NBC dominated through the mid-1970s, when ABC, traditionally regarded as a poor third, rose to the top of the ratings, largely because of shrewd scheduling.

PUBLIC BROADCASTING

A Carnegie Commission report in 1967 recommended the creation of a fourth, noncommercial, public television network built around the educational nonprofit stations already in operation throughout the United States (see television, noncommercial). Congress created the Public Broadcasting System that year. Unlike commercial networks, which are centered in New York and Los Angeles, PBS's key stations, many of which produce programs that are shown throughout the network, are spread across the country. PBS comprises more than 300 stations, more than any commercial network. Some of the most praised programs on PBS, such as the dramatic series *Upstairs, Downstairs* (1971), have been imports from Britain, which has long had a reputation for producing high-quality television. Perhaps the most influential of PBS's original contributions to American television were the educational program for preschoolers, *Sesame Street*, which first appeared in 1969 and is still a popular program.

4.3. Social Networking

Social networking has become an everyday, mainstream way to use the internet. Social networking refers to the use of social media websites and apps, such as *Facebook, Instagram,* and *Twitter*, to connect with family, friends, and people who share your interests.

Social networking is commonplace throughout the world, especially with young people, but not everyone understands exactly what it means. Here's a simple breakdown of social networking's uses, components, and common terms.

If you're participating in social networking, it means ***you're using social media sites, also known as social networks, to connect to others.*** Some of the most popular social media sites are *Facebook, Twitter, Instagram, Snapchat, LinkedIn,* and *Pinterest*.

While various social media sites attract certain types of users, *Facebook* is a good example of a general social network. When you join *Facebook*, you may know some other people who use the site and add them as friends. As you use the platform more, you may add friends who share your interests or discover people you know and add them, as well. ***Other people may find you on Facebook and seek to connect with you.***

The more you interact with a social media site like *Facebook*, the more your network of friends and interests will grow. ***It's similar to networking in real life, for example at a business conference.*** The more you interact with other people and discover common friends and interests, the wider your circle becomes. Social networking sites and apps each have unique features and points of view, but most have common elements. Whether you're starting out with *Facebook Twitter*, or a new site, you'll encounter these terms.

Your Public Profile

Your profile contains basic information about you, including ***a photo, short bio, the town where you live, and sometimes more personal information, such as your birthday, where you went to college, and what your interests are.*** You can usually make your profile as personal or vague as you're comfortable with.

Social networks dedicated to a specific theme, such as music or movies, might need you to supply more information about the topic. For example, dating websites are social networks that focus on making love matches, so you need to be clear about who you are and what you're looking for so you can find a compatible person.

Friends and Followers

Friends and followers are the heart and soul of social networking, adding the social component. Friends and followers are the people you allow to access your profile. They are able to see any photos and posts you make and interact with you via comments and "***likes***". You can also see and interact with their posts.

Some people enjoy getting as many friends and followers as possible, while others prefer a smaller, more intimate group of friends and followers to interact with. Some people even ***set their profile to "public",*** meaning anyone who wants to can follow them or become friends with them. This is often used as a marketing tool.

Home Feed

Most social networking sites have some kind of home page you see when you log on. This usually displays a feed showing updates from friends. Scrolling through your home feed gets you caught up quickly on the activities, thoughts, and news friends want to share.

Likes, Comments, and Shares

Getting and giving feedback is a huge component of social networking. To indicate that a friend or follower has read and appreciated a post, most sites have some kind of ***"like" button,*** perhaps a heart or a thumbs up. *Facebook* has an array of icons you can use to express your reactions to a post, such as sadness, surprise, or love. You're expressing yourself without having to say anything specific.

Most social media sites support comments on posts. Since social networking thrives on interaction, comments are an important element. Whether you comment that someone's baby is cute or make a pointed political observation, comments create conversations and boost the synergy of social networking.

Groups

Some social networking sites have a "*group*" element that helps users find people with similar interests or engage in discussions on certain topics. A group can be anything from ***"Johnson High Class of '98" and "People Who Like Books" to "Doors Fans."***

Social networking groups are both a way to connect with like-minded people as well as a way to identify your interests.

Hashtags

A hashtag is a word or keyword phrase that's preceded by the # symbol, which is called a hash or the *pound sign*. Users put a hashtag by a name or phrase to help others who may be interested in it to find

it when they search for a keyword or particular hashtag. For example, if you post a picture of your baby's cute face and add #babysmiles, people can find it, along with other posts with the same hashtag, if they search for that phrase. Hashtags help to draw attention to your posts and encourage interaction.

Tagging

Tagging is another common element of social networking sites, particularly *Facebook* and *Instagram*. If you post a photo of several people, you can identify another person in the photo by tagging them, usually by clicking on the picture and adding their name. ***Tagging is a way of creating more interaction for your posts.***

All in all, social networking can be enjoyable and entertaining. It's a great way to stay in touch with friends and family and can be an effective promotional tool for businesses, artists, or anyone in need of some exposure.

Social networking lets us reach out to other people with ***similar interests, such as books, television, video games, or movies.*** It can be a great tool for companionship and interaction.

Social networking is for both young and old people, with social media sites that cater to everything from general interests to specific hobbies. There are niche social networks that focus on a specific theme or style of posting.

The top social network sites appeal to a wide variety of users. Try a site that appeals to you and see if it's a good fit. ***You can always leave and try something else.***

Библиографический список

1. *Апресян Ю.Д.*, Цинман Л.Л. Перифразирование на компьютере // Семиотика и информатика. 2022. №. 36. С. 177–202.

2. *Нечаева Н.В.*, Светова С.Ю. Постредактирование машинного перевода как актуальное направление подготовки переводчиков в вузах // Вопросы методики преподавания в вузе. 2018. Т. 7. № 25. С. 64–72.

3. *Прикладная* и компьютерная лингвистика / под ред. И.С. Николаева, О.В. Митрениной, Т.М. Ландо. Изд. 2-е. М.: ЛЕНАНД, 2017. 320 с.

4. *Нуриев В.А.*, Егорова А.Ю. Методы оценки качества машинного перевода: современное состояние // Информатика и ее применения. 2021. Т. 15. №. 2. С. 104–111.

5. *Шереметьева С.О*. Информационные технологии в помощь переводчику: учеб. пособие / С.О. Шереметьева, П.Г. Осминин. Челябинск: ИЦ ЮУрГУ, 2016. Ч. 3. 43 с.

6. *Bahdanau D.*, Cho K., Bengio Y. Neural machine translation by jointly learning to align and translate // arXiv preprint arXiv:1409.0473. 2014.

7. *Frederic Chaume*, Audiovisual Translation: dubbing. 2012. 201 p.

8. *Hutchins W.J*. Machine translation: History and general principles // The encyclopedia of languages and linguistics. 1994. T. 5. P. 2322–2332.

9. *Hutchins W.J*. The Georgetown-IBM experiment demonstrated in January 1954 // Conference of the Association for Machine Translation in the Americas. Springer, Berlin, Heidelberg, 2004. P. 102–114.

10. *Hutchins W.J.*, Somers, H.L. An introduction for machine translation // Computer. 1992. Vol. 25. P. 118–119.

11. *International* Standard ISO 17100:2015. Translation Services – Requirements for translation services. 2015. 19 p.

12. *International* Standard ISO 18587:2017. Translation Services – Post-editing of machine translation output – Requirements. 2017. 15 p.

13. *Jorge D.*, Remael A. Audiovisual translation: subtitling. Routledge, 2007. 284 p.

14. *Russian* Timed Text Style Guide [Электронный ресурс]. URL: https://partnerhelp.netflixstudios.com/hc/en-us/articles/215346638-Russian-Timed-Text-Style-Guide (дата обращения: 20.11.2023).

15. *Timed* Text Style Guide: General Requirements [Электронный ресурс]. URL: https://partnerhelp.netflixstudios.com/hc/en-us/articles/215758617-Timed-Text-Style-Guide-General-Requirements (дата обращения: 20.11.2023).

О Г Л А В Л Е Н И Е

Мамаев Иван Дмитриевич, Шамова Дарья Михайловна

Приложения для автоматизации переводческого процесса

Редактор *А.А. Баутдинова*
Корректор *Л.А. Петрова*
Компьютерная верстка: *Н.А. Андреева*
Подписано в печать 14.03.2024. Формат 60×84/8. Бумага документная.
Печать цифровая. Усл. печ. л. 4. Тираж 100 экз. Заказ № 42.
Издательство БГТУ «ВОЕНМЕХ» им. Д.Ф. Устинова.
190005, С.-Петербург, 1-я Красноармейская ул., д. 1