Санкт-Петербургский государственный университет

**ХУДЯКОВА Юлия Михайловна**

**Выпускная квалификационная работа**

**Сравнительный анализ результатов машинного перевода**

**(на материале русских и французских текстов)**

Уровень образования: магистратура

Направление 45.03.02 «Лингвистика»

Основная образовательная программа ВМ.5805. «Компьютерная и прикладная лингвистика»

Профиль «Компьютерная лингвистика»

Научный руководитель:

доцент, Кафедра математической лингвистики,

Хохлова Мария Владимировна

Рецензент:

доцент, Высшая школа лингводидактики и перевода, Гуманитарный институт, Федеральное государственное автономное образовательное учреждение высшего образования «Санкт- Петербургский политехнический университет Петра Великого»

Коган Мария Самуиловна

Санкт-Петербург

2021

Аннотация

Стирая языковые барьеры и расширяя горизонты коммуникации, технологии машинного перевода становятся неотъемлемой частью нашей повседневной жизни. Результаты автоматического перевода представляют собой все более связные и осмысленные тексты. Для поддержания тенденции популяризации компьютерного перевода проводятся исследования, направленные на оценку его качества и выявление ошибок перевода с целью их последующего исправления. В данной работе особое внимание уделяется сравнению результатов работы нескольких систем машинного перевода.

С помощью детального разбора и классификации расхождений выявляются особенности функционирования каждой из систем. В качестве материала исследования используются русские и французские тексты и их ручные переводы, обеспечивающие возможность проведения анализа в обоих направлениях перевода.

*Ключевые слова*: машинный перевод, сравнительный анализ, системы машинного перевода, классификация ошибок, качество перевода, автоматическая оценка качества, читабельность текста, постредактирование, русский язык, французский язык.

By removing language barriers and expanding the horizons of human communication machine translation technologies are becoming an integral part of our daily lives. The results of automatic translation are becoming more coherent and meaningful texts. In order to maintain the trend of popularization of computer translation many researches are conducted to assess its quality and identify translation errors for their further correction. In this paper special attention is paid to the comparison of the results of several machine translation systems.

Using a detailed analysis and classification of inaccuracies the features of the functioning of each of the systems are revealed. Russian and French texts and their manual translations are used as research material, which allows us to analyze them in both directions of translation.

*Keywords:* machine translation, comparative analysis, machine translation systems, error classification, translation quality, automatic evaluation, text readability, post-editing, Russian, French.

***Оглавление***

[***Введение*** 4](#_Toc73552220)

[***Глава 1*** 7](#_Toc73552221)

[***1.1*** ***Перевод как вид языковой деятельности*** 7](#_Toc73552222)

[***1.2*** ***История развития машинного перевода*** 10](#_Toc73552223)

[***1.3*** ***Обзор существующих систем машинного перевода и их характеристик*** 13](#_Toc73552224)

[***1.4*** ***Особенности перевода языковой пары «русский < > французский»*** 17](#_Toc73552225)

[***1.5*** ***Выводы по Главе 1*** 20](#_Toc73552226)

[***Глава 2*** 22](#_Toc73552227)

[***2.1 Подходы к типологии ошибок машинного перевода*** 22](#_Toc73552228)

[***2.2 Автоматическая оценка качества перевода*** 24](#_Toc73552229)

[***2.3 Автоматическая оценка читабельности перевода*** 27](#_Toc73552230)

[***2.4 Выводы по Главе 2*** 29](#_Toc73552231)

[***Глава 3*** 30](#_Toc73552232)

[***3.1 Цель и задачи эксперимента*** 30](#_Toc73552233)

[***3.2 Составление классификации ошибок*** 30](#_Toc73552234)

[***3.3 Вычисление автоматической оценки качества результатов перевода*** 47](#_Toc73552235)

[***3.4 Вычисление автоматической оценки читабельности результатов перевода*** 49](#_Toc73552236)

[***3.5 Выводы по Главе 3*** 50](#_Toc73552237)

[***Заключение*** 51](#_Toc73552238)

[***Список использованных источников*** 53](#_Toc73552239)

[***Приложение 1*** 57](#_Toc73552240)

[***Приложение 2*** 58](#_Toc73552241)

[***Приложение 3*** 59](#_Toc73552242)

# ***Введение***

Развитие машинного перевода неразрывно связано с увеличением количества переводимой информации. Постоянная необходимость совершенствования качества работы систем машинного перевода свидетельствует об их высокой востребованности. В современном мире становится важно не только обеспечение возможности межъязыковой коммуникации и обмена информацией вне языковых барьеров, но и скорость и эффективность этих процессов. Вследствие технического прогресса машинный перевод находит свое место и в области трудовой деятельности человека: все большую популярность обретает профессия постредактора – специалиста, занимающегося доработкой автоматически переведенного текста [Koponen 2016].

Интерес к системам машинного перевода в профессиональном контексте подтверждает необходимость исследования качества их работы. Одним из методов анализа систем является сопоставление результатов их работы с целью последующего разбора найденных ошибок и определения наиболее совершенного на момент проведения исследования инструмента для постредактирования. В сравнении становится возможно увидеть то, как по-разному функционируют системы, и как это влияет на генерируемый текст.

***Цель*** этого исследования заключается в сопоставительном анализе работы систем машинного перевода для оценки их качества и составления классификации выявленных ошибок. Для достижения данной цели были сформулированы следующие теоретические и практические ***задачи***:

1. Изучение и обобщение источников информации по теме машинного перевода, системах перевода и их особенностях.
2. Определение понятия «качество перевода», анализ существующих подходов к оценке качества перевода.
3. Отбор материала для сравнения результатов перевода.
4. Выявление несоответствий в полученных текстах переводов по сравнению с эталонным образцом.
5. Осуществление детального анализа найденных примеров неточностей в переводах и создание классификации ошибок.
6. Произведение автоматической оценки качества перевода и читабельности сгенерированного текста.
7. Интерпретация и оценка результатов исследования.

***Актуальность*** темы исследования подтверждается тем, что автоматический перевод является одной из наиболее важных задач компьютерной лингвистики и играет существенную роль в повседневной жизни информационного общества в условиях глобализации [Карцева, Маргарян, Гурова 2016]. Необходимость в постоянном улучшении скорости и производительности систем машинного перевода определяет темп их развития: за несколько лет своего существования нейроперевод опередил результаты, достигнутые статистическими системами, на разработку которых было потрачено более двух десятилетий [Худяков 2019]. В настоящий момент использование искусственных нейронных сетей в системах машинного перевода обретает все большую популярность. В этом исследовании продемонстрированы результаты работы сервисов компаний Google, Яндекс и PROMT. Каждый из представленных онлайн-переводчиков обладает высокой производительностью, основывается на современных технологиях, в том числе и на нейросетях, поддерживает русский и французский языки и широко известен среди пользователей Интернета [51].

Таким образом, ***объект*** исследования– это тексты, сгенерированные системами МП Google Translate, PROMT.One и Яндекс – Переводчик в процессе перевода исходных текстов.

В качестве ***предмета*** исследования выступают различия в результатах работы рассматриваемых систем машинного перевода, ошибки и неточности, обнаруженные в сравнении.

Тексты на русском и французском языках и их ручные переводы, на ***материале*** которых базируется исследование, были предоставлены компанией ООО «ЭГО Транслейтинг».

***Практическая значимость*** работы состоит в идентификации основных проблем функционирования популярных и востребованных инструментов машинного перевода. Результаты исследования могут быть применены в лекциях и практических заданиях обучающего курса по постредактированию, тестах для проверки компетентности кандидатов на должность постредактора.

Выпускная квалификационная работа состоит из введения, трех глав, заключения, списка источников информации и трех приложений. В первой главе рассматривается понятие машинного перевода, история создания и развития данного научного направления, приводится обзор систем машинного перевода и описание сервисов, задействованных в исследовании, а также анализируются особенности направлений перевода «русский язык – французский язык» и «французский язык – русский язык». Вторая глава посвящена методам оценки переводов. Рассматриваются принципы составления классификации ошибок перевода, анализируются существующие классификации. Разбираются способы автоматической оценки переводов с помощью статистических метрик. В качестве эксперимента в исследовании также упоминается автоматическое вычисление читабельности результирующего текста. Третья глава исследования содержит информацию о практической части работы: этапах предобработки текстов, создании собственной классификации ошибок переводов, результатах вычисления автоматической оценки качества и читабельности переводов. В приложениях приводятся программные решения практического этапа исследования.

# ***Глава 1***

## ***Перевод как вид языковой деятельности***

Определение сущности перевода на протяжении многих лет является одним из важных теоретических аспектов лингвистики. Под переводом принято понимать интерпретацию одного вербального знака в терминах другого. Различают три основных процесса: переименование, трансмутацию и перевод. Переименование представляет собой внутриязыковой перевод, когда лингвистический знак объясняется через синоним или перефразирование. Трансмутация означает межсемиотический перевод, в рамках которого элемент одной знаковой системы выражается с помощью средств другой. Понимание же перевода как межъязыкового явления подразумевает перекодирование исходной информации в вербальных знаках другого языка [Якобсон 1978]. Важно отметить, что в этом случае перевод представляет собой средство межкультурной коммуникации.

Основной задачей межъязыкового перевода становится обеспечение возможности обмена информацией между носителями разных языков. При подобной трансформации текста должны быть учтены особенности культурного кода. От общности культурных компонентов языков, участвующих в процессе перевода, напрямую зависит эквивалентность результата перевода [Владова 2011]. Некоторые ученые объединяют понятия эквивалентности и адекватности перевода, однако в большинстве случаев в исследованиях подчеркивается их разница. Адекватность гарантирует такое качество перевода, которое допустимо и необходимо в рамках конкретной ситуации. В то же время эквивалентность перевода подразумевает максимальное смысловое совпадение между исходным текстом и конечным, которого в рамках этой конкретной ситуации можно было достичь [Бархударов 1975].

Язык, на котором написан оригинальный текст, принято называть исходным языком (ИЯ), а язык вторичного текста — языком перевода (ЯП) [Комиссаров 1990].

Генерация нового текста на ЯП в результате перевода представляет собой синтез текста. Формальное описание процесса синтеза текста приводит Ю. Д. Апресян, определяя синтез как следующий шаг после понимания. Понимание текста также происходит в несколько этапов: морфологический анализ предшествует синтаксическому и семантическому разборам (см. Рисунок 1). Производство же текстов представляет собой обратный процесс: от семантического анализа - к синтаксическому и морфологическому. Последовательное выполнение каждого из этапов обработки текста предполагает сохранение смыслового содержания оригинального фрагмента и точность и полноту синтезируемого фрагмента [Апресян 1992].

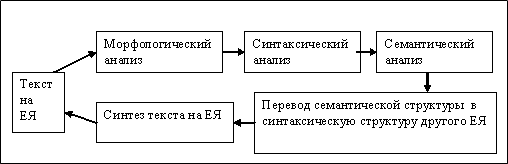


Рисунок 1. Структура процесса анализа текста

Центральным объектом данного исследования становится машинный перевод (МП). Понятие МП формулируется добавлением к определению межъязыкового перевода условия выполнения перевода компьютером. Аналогично ручному переводу, это перекодирование значения исходного текста в знаках другого языка. В широком смысле слова, машинный перевод — это междисциплинарная научная область, связанная как с переводоведением и лингвистикой, так и с информационными технологиями: она использует словари и грамматические правила наряду с программными средствами и компьютерным оборудованием. Основным преимуществом МП является время работы: результат перевода текста любого объема можно получить за считаные секунды. Быстрота и дешевизна перевода становятся определяющими факторами в востребованности МП и подтверждении необходимости его развития с целью улучшения качества [Белобородова, Бодулева 2016]. Однако традиционно качество результатов машинного перевода было в лучшем случае достаточным для того, чтобы кратко доносить суть первичного текста для пользователей, не владеющих исходным языком. Выигрывая в скорости, системы машинного перевода зачастую предлагают результат недостаточного качества, который требует доработок и редактуры со стороны человека. Так появляется понятие автоматизированного перевода (АП) – технологии, при которой результаты машинного перевода впоследствии обрабатываются и корректируются человеком [Zong 2018]. Автоматизированный перевод становится неотъемлемой частью профессиональной переводческой среды и служит для оптимизации рабочего процесса, а также снижения количества трудозатрат и стоимости услуг перевода без потери качества. Являясь следствием глобализации, растущие потребности в коммуникации и обмене информацией приводят к появлению огромного количества машинных переводов текстов для личного и коммерческого использования. Автоматизированный перевод сочетает качество человеческого перевода и эффективности компьютерных технологий и является основой такой профессиональной отрасли, как постредактирование. Процесс постредактирования характеризуется коррекцией результатов машинного перевода для обеспечения надлежащего уровня качества, который заранее согласовывается между клиентом и постредактором [Carl 2015].

Тем не менее, в связи с разницей в качестве результатов работы существующих систем МП и их зависимостью от языка оригинала и языка перевода, все больше и больше переводчиков сталкиваются с увеличением рабочей нагрузки в задачах постредактирования. Таким образом, сопоставительный анализ результатов машинного перевода разных систем направлен на поиск самого эффективного решения задач постредактирования.

## ***История развития машинного перевода***

Для того чтобы перейти к анализу существующих достижений в области компьютерного перевода, следует обратиться к его истории. Обзор исторических событий, сопровождавших развитие МП, позволяет сопоставить данные об используемых в разные временные периоды технологиях, проследить особенности их функционирования и по-новому взглянуть на те разработки и инструменты, которые сейчас находятся в свободном доступе и открыты для изучения. Идея о делегировании машине перевода текстов на естественном языке появилась еще в 19-м веке, но воплотилась в реальность только в конце 20-го [Peng 2018]. Изобретатель первой аналитической вычислительной машины Чарльз Бэббидж упоминал о возможности автоматического перевода разговорной речи [Bromley 1982]. Несмотря на то что это предположение не было реализовано в момент своего зарождения, его можно считать «отправной точкой» в череде событий, повлекших за собой достижения машинного перевода, доступные в настоящий момент.

Исследователи выделяют несколько основных этапов в истории становления машинного перевода [Дроздова 2015, Кулагина 1991, Chéragui 2012, Hutchins 1995]. В этой главе приведена обобщенная и систематизированная информация о временных границах каждого из этапов и об открытиях, сделанных в этот период в сферах МП и АП.

* Этап 1 – **«Начало» (1949-1960):** первые идеи о концепции машинного перевода и появление термина «computer translation» в «Меморандуме» Уоррена Уивера (1949 г.), первая конференция, посвященная машинному переводу, в Массачусетсе (1952 г.), «Джорджтаунский эксперимент» - изобретение первого автоматического переводчика в университете Джорджтауна в соавторстве с IBM и представление первой системы МП – IBM Mark II (1954 г.);

Одним из важнейших шагов первого этапа развития МП стал совместный эксперимент американской корпорации IBM и университета Джорджтауна. Суть эксперимента заключалась в попытке перевода на английский язык 60-ти русских предложений, объединенных тематикой нефтехимии. Выбор направления перевода был обусловлен историческим контекстом начала Холодной войны и необходимостью интерпретации огромных потоков информации на русском языке. В то же время в СССР в Институте точной механики и вычислительной техники был проведен обратный эксперимент по переводу с английского языка [Кулагина 1991]. Результаты экспериментов казались вполне успешными, но дальнейшие исследования более поздних годов подтвердили то, что проблема машинного перевода была сильно недооценена.

* Этап 2 – **«Первые разочарования» (1960-1970):** появление термина «компьютерная лингвистика» в лекциях Дэвида Хэйса (1961 г.), учреждение комитета ALPAC (Automatic Language Processing Advisory Committee) при американском правительстве для изучения перспектив и возможностей машинного перевода (1964 г.), доклад ALPAC о нерентабельности МП (1966 г.);

Разработки машинного перевода, относимые ко второму периоду, имели в своей основе технологию пословного перевода. При выборе эквивалента игнорировалось условие смысловой многозначности слов, что негативно отражалось на качестве перевода. Не учитывались и синтаксические зависимости между словами. Второй период развития МП характеризуется резким спадом в прогрессе научной отрасли: выступление Консультативного комитета по автоматической языковой обработке (ALPAC) в 1966-м году подтвердило несостоятельность машинного перевода в его текущем состоянии и повлекло за собой сокращение финансирования разработок этой области.

* Этап 3 – **«Новые начинания» (1970-1990)**: разработка систем SYSTRAN1 и FORTRAN (1970 г.), создание системы TAUM-MÉTÉO проектом TAUM, занимающимся машинным переводом в университете Монреаля (1976 г.), создание системы ATLAS2, основывающейся на правилах, корпорацией Fujitsu в Японии (1978 г.), PIVOT от NEC c использованием интерлингвы (1983 г.), появление системы TRADOS, основывающейся на технологии «translation memory» (1984 г.), выход на рынок англо-японской системы DUET с функцией оптического распознавания символов (1988 г.);

В начале 70-х годов в связи со скачком в развитии компьютерных технологий сфера машинного перевода переживает «второе рождение». После неутешительных выводов 60-х годов отношение к машинному переводу меняется — это больше не полностью автономный инструмент. Восприятие автоматического перевода как помощника переводчика влияет на вектор развития систем МП: в разработках применяются технологии морфологического и синтаксического анализа, появляются системы, в основе которых лежат лингвистические правила. К примеру, система SYSTRAN, появившаяся в 1970-м году, сменяет лидера отрасли - IBM Mark II и успешно распространяется не только в США, но и в Европе [Кенджаев 2014]. Большое количество проектов, основанных в 70-80-е гг., в дальнейшем развились в полноценные коммерческие системы. Важнейшим достижением этого этапа развития МП стала технология памяти перевода – «translation memory». Аналогично человеческой памяти, в ее основе находится принцип сохранения и накопления информации: из исходных текстов формируется база данных, примеры из которой при похожести на искомый сегмент предлагаются переводчику с указанием процентного совпадения. Данная технология применяется и в современных системах МП.

* Этап 4 – **«Интернет и перевод» (1990-2010)**: Выход на рынок отечественных разработчиков – презентация первой коммерческой системы машинного перевода PROMT (1990 г.), создание системы перевода ALPH, основанной на примерах (1999 г.), появление первого вебсайта с онлайн-переводом GOOGLE (2005 г.), появление первой гибридной системы METIS-II (2007 г.);

С появлением Интернета и многократным увеличением переводимого контента отрасль машинного перевода стала развиваться в ускоренном темпе. Для удовлетворения потребностей пользователей потребовались новые системы МП с разнообразной архитектурой. Рост популярности машинного перевода обеспечил стабильное место в коммерческой среде компаниям, занимающимся разработкой систем МП и в настоящий момент: IBM, SYSTRAN, Lernout & Hauspie, Language Engineering Corporation, Deep Genomics, Transparent Language, Unbabel, Nova Incorporated, Trident Software, Atril, TRADOS, Linguatec, PROMT, Caterpillar Co., LingoWare, Ata Software и др. [Кольцова, Кольцов 2019].

* Этап 5 – **«Искусственный интеллект» (с 2010 г.):** появление идей о внедрении технологии deep learning, первая научная статья об использовании нейронных сетей в машинном переводе (2014 г.), демонстрация NMT-системы в публичном конкурсе машинного перевода OpenMT (2015 г.), использование нейросети в Google Translate (2016 г.);

Технологии, основанные на глубинных нейронных сетях, бывшие реформаторами в отрасли машинного перевода, сейчас стали частью повседневной жизни людей, а эффективность современных систем МП достигла небывалых показателей [Мохов 2019]. Анализ существующих сервисов и актуальности поддерживаемых ими технологий позволит еще более подробно охарактеризовать положение машинного перевода в настоящий момент.

## ***Обзор существующих систем машинного перевода и их характеристик***

На сегодняшний день исследователи выделяют три основных подхода к машинному переводу: перевод на основе правил (*rule-based machine translation*), статистический машинный перевод (*statistical machine translation*) и гибридный перевод (*hybrid machine translation*) [Митренина 2017]. Чтобы подробнее изучить принципы работы систем машинного перевода, основывающихся на правилах, следует обратиться к иллюстрации пирамиды французского математика Бернарда Вакуа (см. Рисунок 2).

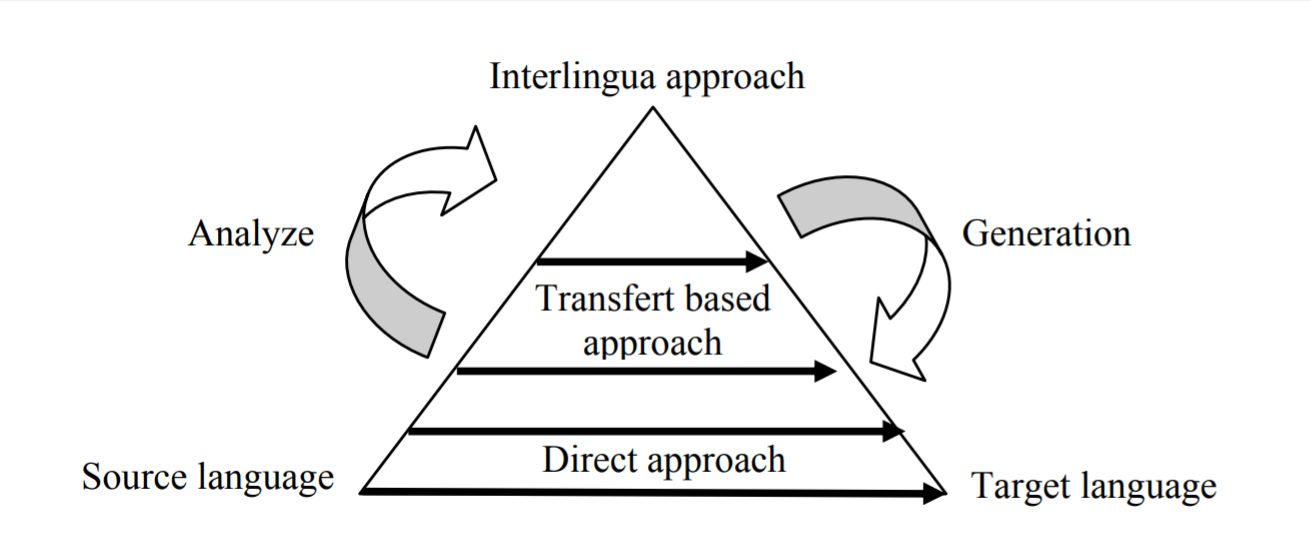


Рисунок 2. Пирамида Бернарда Вакуа

На данном схематическом изображении представлены три возможных варианта межъязыкового перевода, реализованных в правиловых системах: пословный перевод (*direct approach*), трансферный перевод (*transfer based approach*) и интерлингвистический перевод *(interlingua approach*).

На первом уровне пирамиды расположены системы пословного перевода, и, если обратиться к истории направления, изложенной в 1.2, они являются самыми первыми и самыми примитивными системами МП. Между ИЯ и ЯП отсутствуют дополнительные шаги: каждому слову исходного текста находится соответствие на языке перевода. Такие системы не учитывают синтаксические связи и зависимости между словами, поэтому результат хоть и сохраняет минимальное смысловое значение первичного фрагмента, обладает недостаточным для адекватного понимания текста качеством [Митренина 2017].

Функционирование трансферных систем схоже с этапами работы человека-переводчика. На этапе анализа текст на ИЯ рассматривается с точки зрения синтаксиса и морфологии. Иногда трансферные системы включают и семантический разбор. Трансфер подразумевает «преобразование структуры входного предложения с учетом требований языка перевода» [Дроздова 2015]. Конечным этапом работы систем является синтез результата.

Наконец, системы, использующие интерлингву – язык-посредник, трансформируют первичный текст в некое абстрактное представление [Гриф, Тимофеева 2012]. Реализация интерлингвистической системы представляет собой сложный многоступенчатый процесс. Существует предположение о том, что главным преимуществом интерлингвы является возможность легкого расширения базы знаний и добавления в систему новых языковых данных, однако на данный момент в сфере машинного перевода не представлено ни одного полноценно функционирующего вспомогательного языка [Митренина 2017].

Статистический машинный перевод – это подход, опирающийся на методы машинного обучения. В его основе находится статистический анализ межъязыковых соответствий: при создании системы машина получает лингвистическую базу предложений и их переводов. В системе хранится информация о контексте, частоте совместной встречаемости слов и их месте в составе предложения. Запоминание частот различных цепочек слов позволит компьютеру выбрать самый вероятный вариант перевода того или иного слова [Митренина 2017]. Кольцова Д. А. и Кольцов С. В. отмечают возможность самообучения таких систем, быстроту настройки и отладки и легкость добавления новых переводных направлений.

Системы гибридного перевода носят такое название в связи с тем, что они сочетают в себе архитектуры, описанные ранее, и, согласно ожиданиям, объединяют в себе их преимущества [Андреева 2013]. До недавнего времени гибридные системы считались совокупностью правил и статистического анализа, однако с развитием искусственных нейронных сетей, они стали включать и нейронный машинный перевод. Об устройстве нейросистемы МП пишет Р. Г. Мифтахова: «Искусственная нейросеть состоит из уровней искусственных нейронов, которые связаны, так называемым, весом, т.е. параметрами сети». Важной особенностью искусственной сети становится ее способность к самостоятельному изменению собственных параметров во время обучения. Синтезированный текст на языке перевода сравнивается с ожидаемым эталоном, после чего исправленный вариант возвращается в систему и меняет связи между элементами сети в соответствии с внесенными исправлениями.

В данном исследовании для сравнения результатов машинного перевода и идентификации распространенных ошибок используются популярные онлайн-сервисы компаний Google [48], Yandex [45] и PROMT [50], находящиеся в свободном доступе. Для большей наглядности параметры рассматриваемых сервисов представлены в Таблице 1.

Данные о функционале систем были получены из их документации. При сравнении указанных в Таблице 1 параметров становится возможно сделать выводы о преимуществах каждой из систем: сильной стороной Google Translate является большое количество языковых пар, а система PROMT выделяется доступностью множества сервисов, полезных в автоматизированном переводе и постредактировании. Яндекс. Переводчик, будучи относительно новой разработкой, уже реализует 98 направлений перевода и, как и GT и PROMT, основывается на самых современных технологиях МП – сочетании статистического подхода и применения нейросетей.

|  |  |  |  |
| --- | --- | --- | --- |
| Параметр | Google Translate | Яндекс. Переводчик | PROMT. One |
| Используемые технологии | Гибридный МП:  с 2006-го года включает статистический подход, с 2016 г. - нейроперевод | Гибридный МП: с 2011-го года включает статистический подход, с 2017-го года - нейроперевод | Гибридный МП: с 2000-го года включает статистический подход, с 2015 г. - нейроперевод |
| Языки перевода | 108 языков | 98 языков | 22 языка |
| Объем вводимого текста | 5000 символов на странице | 10000 символов | 3000 символов, после регистрации – 10000 символов |
| Автоматическое определение ИЯ | + | + | + |
| Исправление результата | + | + | - |
| Выбор тематики исходного текста | - | - | +  (Автомобили, деловая переписка, компьютеры, медицина, частная переписка) |
| Дополнительные инструменты | Гендерная специфика перевода, голосовой перевод, фотоперевод | Голосовой перевод, фотоперевод | Разделы «Словарь», «Спряжение и склонение», «Контексты», голосовой перевод, фотоперевод |

Таблица 1. Сопоставление функционала онлайн-сервисов компаний Google, Yandex и PROMT, задействованных в практической части исследования.

## ***Особенности перевода языковой пары «русский < > французский»***

Одним из важных шагов работы становится рассмотрение особенностей перевода исследуемой пары языков. Проведя анализ их различий и систематизировав информацию о часто встречаемых трудностях перевода, не зависящих от характеристик конкретных переводимых текстов, будет значительно легче перейти к этапу анализа существующих классификаций ошибок машинного перевода.

Согласно Б. Д. Добровольскому, основные проблемы перевода можно разделить на языковые и внеязыковые. К внеязыковым трудностям относятся отсутствие лексических эквивалентов, социокультурные коннотации и контекст, адаптация имен собственных. В свою очередь, языковые проблемы делятся на «очевидные», то есть лежащие на поверхности и периодически видимые даже визуально при сопоставлении текстов, и «скрытые». К очевидным языковым проблемам относятся различия в грамматическом строе языка, несоответствия орфографии и фонетики. Среди скрытых лингвистических проблем выделяются особенности стилистики, отсутствие взаимозаменяемости синонимов, лексико-фразеологическую сочетаемость и др. [Добровольский 2009]. Аналогичный вариант классификации трудностей перевода предлагает в своей монографии Томилова А. И., разделяя проблемы подхода к переводу на «лингвистические» и «экстралингвистические». К экстралингвистическим трудностям она относит не только социокультурные проблемы, но и слабую языковую компетенцию переводчика и недостаточность знаний [Томилова 2018].

А. Е. Александрова, анализируя взаимную переводимость русского, французского и английского на примере перевода каламбуров, делает вывод о существенном влиянии лексического сходства языков. Так, в паре «русский – французский» обнаружился сравнительно небольшой процент межъязыковых соответствий из-за отсутствия исторического родства и малого взаимовлияния культур в сравнении с английским языком [Александрова 2019].

Существенный вклад в развитие направления перевода вносит советский лингвист Владимир Григорьевич Гак, автор множества книг, посвященных теории и практике переводоведения. Анализируя перевод эквивалентов в русском и французском языках, он выделяет три модели поведения: подбор формальных эквивалентов (использование аналогичных языковых форм), подбор смысловых эквивалентов (разные способы выражения общего значения) и подбор ситуационных эквивалентов (поиск слов, являющихся эквивалентами только в конкретном контексте) [Гак 1977]. В данном исследовании будет проанализировано качество подбора эквивалентов на уровне машинного перевода, и при выявлении использования компьютером некорректных синонимов, пример будет помещаться в специальный раздел классификации ошибок.

Среди прочих скрытых языковых проблем В. Г. Гак выделяет:

* Учет широты значения слова, упоминая, что многим понятиям французского языка в связи с их широкой семантикой соответствуют значения ряда русских слов;

*Пример*: глагол avoir сочетается с референтом через значение владения и «сохраняет во всех своих употреблениях инвариантный характер обозначаемого признака» [Федюк 2009];

* Перевод фразеологических выражений и сочетаемость слов – например, в французском языке зависит от существительного, и некоторые глаголы при полном соответствии значений не могут употребляться в конкретной ситуации;

*Пример:* глаголы «rendre», «accorder», «exercer» в значении «оказать» употребляются с разными существительными: «оказать услугу» переводится «rendre un service», «оказать помощь» — «accorder son aide», «оказать влияние» — «exercer son influence»);

* Разницу в прямом и переносном, самостоятельном и служебном значениях слова;

*Пример:* конструкции с имплицитным отрицанием в русском языке – «нужны мне ваши деньги», «от вас дождешься!» и др. [Люльчева 2013];

* Особенности и расхождения в морфологии и синтаксисе, в частности, из-за разницы в грамматическом строе языков: синтетический строй русского языка сопоставляется аналитическому строю французского;
* Стилистические характеристики слов: нейтральная (нулевая стилистическая характеристика), положительная (выражение радости, счастья, торжественности и др.) или отрицательная (выражение гнева, недовольства, иронии и др.) эмоциональная окраска. Зачастую одно французское слово обладает рядом эквивалентов русского языка с различной стилистической характеристикой, и наоборот (данное направление встречается реже);

*Пример*: суффикс прилагательного, означающий интенсивность цвета,

-atre- может придавать слову отрицательную стилистическую характеристику;

* Грамматические трансформации: важно учитывать обязательность грамматических замен в случае отсутствия искомой грамматической категории в переводящем языке, при использовании в конечном тексте перевода конструкций, не существующих в исходном языке, а также в ситуации различий в использовании стилистических норм в обоих языках;\

*Пример*: частое употребление настоящего времени в повествовании французского языка и редкое – в повествовании русского языка;

* Сохранение смыслового центра предложения. Является распространенной ошибкой в сложных, нагруженных придаточными конструкциях [Гак 1977];

Важно понимать, что трудности перевода языковой пары «русский < > французский» не ограничиваются представленными примерами, но использование их в качестве опоры для анализа существующих классификаций ошибок машинного перевода и составления собственной классификации представляется продуктивным.

## ***Выводы по Главе 1***

Таким образом, в Главе 1 данного исследования были подробно рассмотрены определение и виды перевода как лингвистического явления. Посредством анализа существующих подходов к переводу было сформулировано понятие машинного перевода. Для раскрытия полноты понятия и единообразия обозначений уточняются термины автоматизированного перевода и постредактирования, адекватности и эквивалентности перевода, исходного и переводящего языков. Рассмотрена история развития направления машинного перевода, определены предпосылки состояния сферы МП в настоящий момент. Описаны существующие виды систем машинного перевода, приведена краткая характеристика сервисов компаний Google, Yandex и PROMT, которые участвуют в этом исследовании. Каждая из рассматриваемых систем использует гибридные технологии, сочетающие статистический перевод и использование нейросетей.

Отдельным пунктом обобщены выделенные в переводоведении особенности перевода направлений «русский – французский» и «французский - русский», приведены распространенные типы ошибок перевода.

Систематизация теоретического материала по теории перевода позволит компетентно интерпретировать результаты практической части исследования.

# ***Глава 2***

## ***2.1 Подходы к типологии ошибок машинного перевода***

Цель данного исследования формулируется как сравнение работы систем машинного перевода для языковой пары «русский язык < > французский язык». Одним из классических методов оценки результатов МП является сопоставление конечного текста и эталонного варианта перевода. В результате сравнения обнаруженные несоответствия в результате обобщения и систематизации включаются в состав общей классификации ошибок. Универсального варианта классификации не существует: исследователи могут включать в ее состав малейшие неточности или же заострять внимание только на серьезных искажениях смысла. В этой работе будут рассмотрены несколько существующих вариантов классификаций ошибок МП в рассматриваемом направлении перевода, а также будут выявлены условия, которые впоследствии будут учтены при составлении собственной типологии.

В работе Бунтман, Гончарова и Нуриева приводится пример создания таргетированной таксономии ошибок МП в направлении «русский -> французский». Корпус исследования составил 3000 параллельных контекстов, по 1500 на каждую систему перевода: в работе анализировались сервисы Google и Yandex. В основе классификации находятся два типа ошибок: грамматические и лексические ошибки в фрагменте текста с коннектором и ошибки в переводе самого коннектора. Под термином «коннектор» в исследовании понимается союзное слово или словосочетание. Каждому типу ошибки присваивалось свое сокращенное обозначение, записанное в формате CamelCase, к примеру, для безошибочного фрагмента перевода использовался тег «NoError». Итоговый вариант классификации представлен на Рисунке 3.

Для сравнения в ходе исследования были рассмотрены другие варианты таксономий: для языковой пары «русский - английский» [Томин 2015] и «русский - немецкий» [Кочеткова, Ревина 2017]. Русско-английская классификация составлена линейно: выделено 16 типов ошибок, не объединенных между собой, приведено 22 примера. Классификация для русского и немецкого языка имеет иерархическую структуру, ошибки в ней подразделены по типам: лексические, морфологические и синтаксические. Всего выделено 10 типов ошибок: отсутствие перевода лексической единицы, частичный перевод лексической единицы, потеря смысла фрагмента, несоответствие падежа, числа, наклонения, времени, предлога, степени сравнения, неправильное расположение слов в предложении. Суммарно обнаружено 400 примеров несоответствий.

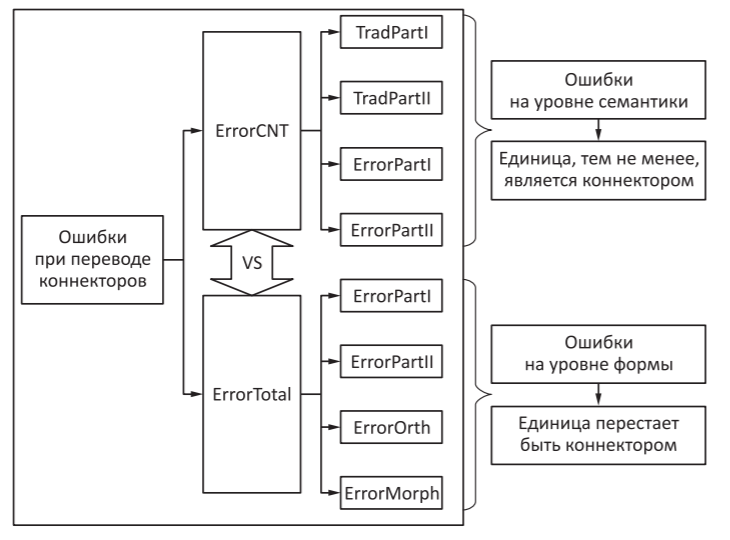


Рисунок 3. Классификация ошибок МП в направлении «русский-французский» в исследовании Н. В. Бунтман, А. А. Гончарова, В. А. Нуриева

Таксономии ошибок МП, созданные при использовании эталонных переводов и участии в процессе профильных специалистов – лингвистов и переводчиков, в большей степени способны повлиять на улучшение качества машинного перевода. Идентификация ошибок в первую очередь ориентирована на выявление проблем устройства и функционирования систем машинного перевода. [Бунтман, Гончаров, Нуриев 2019]. В данном исследовании в дополнение к составлению классификации ошибок перевода русских и французских текстов, будут проведены эксперименты по вычислению автоматической оценки качества перевода и подсчету индекса читабельности результирующего текста с целью проверить применимость такого типа оценки результатов МП в процессе постредактирования.

## ***2.2 Автоматическая оценка качества перевода***

Выделяют два основных подхода к оценке качества машинного перевода: задействование экспертов и использование автоматизированных метрик. Экспертная оценка считается наиболее качественным методом, однако применяется достаточно редко ввиду необходимости участия специалистов, высокой стоимости и больших временных затрат. Среди основных критериев экспертной оценки упоминают точность и естественность. Критерий точности отвечает за полноту передачи смысла оригинального фрагмента, естественность подразумевает под собой соблюдение лексических и грамматических норм переводящего языка. При сопоставительном анализе результатов работы разных систем МП суть экспертной оценки заключается в ранжировании: специалистам необходимо расставить представленные варианты переводов в порядке повышения качества [Рычихин 2019]. Новый вариант экспертной оценки, опирающийся всего лишь на 2 параметра, был предложен в Цюрихском университете. Специалистами оценивалось соответствие исходному тексту (*adequacy*) и плавности (*fluency*) посредством ответов на вопросы «Какой перевод лучше передает смысл исходного текста?» и «Какой перевод грамматически более правилен?». Экспертам предлагалось оценивать как отдельные предложения, так и целые тексты. Несмотря на то, что эксперты предсказуемо сделали выбор в пользу ручного перевода, эксперимент продемонстрировал и несколько полезных для развития МП выводов: если в параметре адекватности МП мог конкурировать с человеческим переводом, то в параметре гладкости однозначно ему уступал. Также ученые заострили внимание на необходимости выхода методик оценки качества МП на более обширный контекстный уровень [Ивтушок 2018].

В этой работе особое внимание уделяется метрикам автоматической оценки качества перевода. Гипотеза исследования заключается в возможности применения данного подхода в работе постредактора: способна ли автоматическая оценка качества перевода продемонстрировать, какую систему МП лучше выбирать при постредактировании текстов конкретной языковой пары? В практической части работы для проверки гипотезы результаты вычисления автоматизированных метрик будут сопоставлены с процентным соотношением ошибок, допущенных каждой из систем.

Важно отметить преимущества использования метрик автоматической оценки. К ним относятся быстрота выполнения, низкая стоимость и минимум контроля со стороны человека. В данной работе будут использоваться метрики BLEU *(BiLingual Evaluation Understudy)*, NIST (от названия университета - *National Institute of Standards and Technology*), METEOR (*Metric for Evaluation of Translation with Explicit ORdering*) и ROUGE (Recall-Oriented Understudy for Gisting Evaluation). Для понимания их различий следует привести более детальную характеристику каждой из метрик.

1. BLEU или «двуязычная оценка дублера» была предложена корпорацией IBM в 2002-м году [Papineneni 2002]. В ней учитывается количество совпадающих в эталонном и синтезированном компьютером текстах перевода n-грамм. После расчета совпадений производится оценка качества перевода по формуле (см. Рисунок 4). Анализ именно n-грамм, а не отдельных слов является большим преимуществом BLEU, но это влияет на ее работоспособность на разных объемах текста.

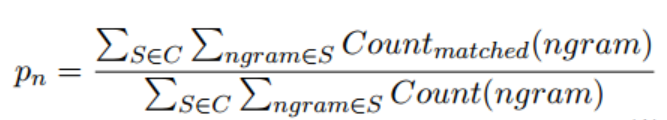


Рисунок 4. Формула вычисления точности n-грамм BLEU

1. NIST – это алгоритм оценки качества машинного перевода, разработанный в американском Национальном институте стандартов и технологий. Его существенное отличие от BLEU заключается в вычислении не только точности, но и информативности n-грамм. Также, в отличии от BLEU, небольшие изменения в объеме исследуемого текста не повлияют на результат работы метрики [Легостина 2019].
2. METEOR – метрика, учитывающая синонимичность и точное сопоставление слов друг другу. Ее разработчиком является Институт языковых технологий Питтсбурга. «В данной метрике принимается во внимание вариация переводов в виде флексии, синонимов и изложения одинакового содержания с помощью других слов» [Кушеева, Гомбожапова, Аюшеева 2019]. Аналогично BLEU, первым шагом METEOR учитывает количество совпадений униграмм в предложениях.

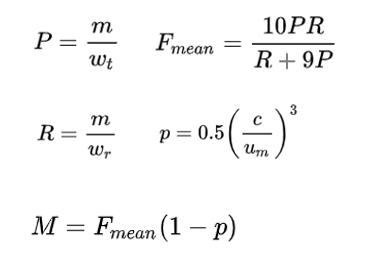


Рисунок 5. Этапы вычисления метрики METEOR [Banerjee, Lavie 2005]

1. ROUGE подсчитывает число перекрытий (n-граммы слов) текста автоматического перевода с «золотым стандартом» перевода, сделанного профессионалом. ROUGE требует довольно простых расчетов, по сравнению с другими рассматриваемыми метриками.

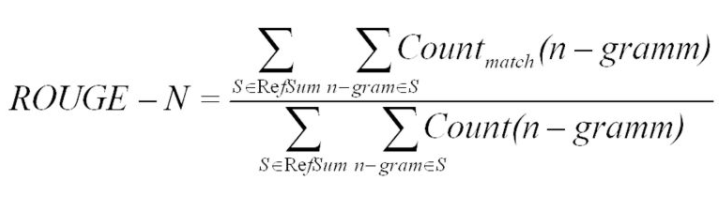


Рисунок 6. Формула вычисления метрики ROUGE [Lin, Chin-Yew 2004]

## ***2.3 Автоматическая оценка читабельности перевода***

Заключительной гипотезой исследования становится возможность применения метрик расчета индекса читабельности текста в постредактировании. Предположение заключается в наличии взаимосвязи между удобочитаемостью текста и легкостью его редактирования. В исследовании предполагается, что показатель читабельности должен отражать количество трудозатрат, которые постредактору необходимо будет приложить для совершенствования синтезированного компьютером перевода.

Ю. В. Рогушина формулирует понятие удобочитаемости как «меры доступности информации, содержащейся в тексте». Признаками сложного для восприятия читателем текста становятся громоздкие синтаксические структуры, большое количество узконаправленной лексики и терминов. Самыми распространенными показателями, определяющими меру читабельности, становятся количество и длина слов и предложений в тексте. В середине 20-го века для единообразия процесса оценки читабельности текста начинают появляться формулы, учитывающие некоторые из этих показателей. Одной из таких формул является автоматический индекс удобочитаемости (*automated readability index, ARI*). Вычисление данной метрики достаточно простое ввиду легкости подсчета элементов, участвующих в ней. Задействованы две переменные: среднее количество букв в слове и среднее количество слов в предложении (см. Рисунок 7).



Рисунок 7. Формула подсчета метрики ARI

Другой популярной формулой оценки удобочитаемости текста считается метрика Флэша-Кинкейда (см. Рисунок 8). Подсчет индекса осуществляется путем вычисления отношений между общим количеством слов, предложений и слогов текста. Подразумевается, что учет количества слогов в метрике позволяет добиться более качественных результатов оценки. Принцип работы индекса F-K аналогичен ARI, и рассчитан на применение в определении возраста целевой аудитории текста.



Рисунок 8. Формула подсчета метрики F-K

Формула SMOG (см. Рисунок 9), как и F-K, является простым статическим уравнением, учитывающим количество слогов в слове. Эта метрика в качестве первой переменной принимает количество слов, в которых 3 и больше слогов, а в качестве второй переменной использует общее количество предложений. Данная формула обычно демонстрирует более высокие показатели, так как предназначена для прогнозирования абсолютного понимания текста [Солнышкина, Гафиятова, Казачкова 2020].

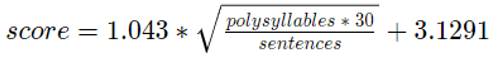


Рисунок 9. Формула подсчета метрики SMOG

Индекс Колман-Лиау (см. Рисунок 10), в отличие от F-K и SMOG, учитывает не количество слогов, а количество символов в слове.



Рисунок 10. Формула подсчета метрики C-L

Метрики оценки читабельности текста вызывают много разногласий в вопросах справедливости их применения. Некоторые ученые, анализируя состоятельность приведенных выше формул, подчеркивают, что данные уравнения действительно можно использовать для оценки читаемости. Однако не следует принимать их результаты за абсолютную истину и оценивать различия в десятых долях оценки как значимые. Также эти уравнения не учитывают нетекстовые элементы, зачастую преобладающие в текстах: заголовки, маркированные списки, уравнения, скрипты и рисунки [Zhou, Jeong, Green 2017].

## ***2.4 Выводы по Главе 2***

Вторая глава исследования посвящена обобщению теоретической информации о методах оценки качества перевода. Рассмотрены три существующие таксономии ошибок МП. Выявлены преимущества и недостатки экспертной оценки качества автоматического перевода.

На основании описанных в разделах 2.2 и 2.3 метрик в практической части исследования будет произведена оценка качества и читабельности результатов перевода каждой из рассматриваемых систем перевода. Расчеты позволят доказать или опровергнуть высказанные гипотеза исследования о применимости данных метрик в автоматизированном переводе и работе постредактора.

# ***Глава 3***

## ***3.1 Цель и задачи эксперимента***

Целью эксперимента данного исследования являлось применение разных подходов к оценке качества машинного перевода в сопоставлении результатов работы рассматриваемых сервисов. Отдельно анализировалась возможность использования данных методов в практике автоматизированного перевода и постредактирования. Исходя из поставленной цели были определены следующие задачи:

1. Сформулировать критерии текстов, которые будут использоваться в работе в качестве оригиналов.
2. Отобрать подходящие фрагменты на русском и французском языках и составить параллельный корпус.
3. Определить набор инструментов, применяемых в программных решениях.
4. Написать скрипты для выделения несовпадений в парах текстов, автоматической оценки качества перевода и читабельности с применением метрик, выделенных в Главе 2.
5. Составить классификацию ошибок МП на материале обработанных текстов.
6. Вычислить и интерпретировать значения, полученные в результате автоматической оценки.
7. Сопоставить результаты таксономии ошибок и оценки качества и читабельности переводов.

## ***3.2 Составление классификации ошибок***

Первым этапом практического исследования стало определение стилистической направленности рассматриваемых текстов. Предпочтение отдавалось официально-деловому стилю в связи с такими факторами, как нейтральный тон изложения, преобладание именных конструкций, однозначность понимания смыслового содержания документа [Замуруева 2010]. Анализ текстов этого стиля гарантирует исследование качества перевода узконаправленной лексики и сохранения смысловых зависимостей в нагруженных синтаксических конструкциях. Важным условием в оценке перевода официальных текстов является сохранение оригинальных дефиниций терминологии. Отсутствие образных средств, употребляемых в переносном смысле, в официально-деловом стиле гарантирует невозможность разночтений и в то же время увеличивает серьезность искажения смысла в случае ошибки. Точность и лаконичность изложения значительно облегчают последующую интерпретацию выявленных несоответствий в переводах. Таким образом, в соответствии с сформулированными критериям, при сотрудничестве компании ООО «ЭГО Транслейтинг» были отобраны тексты, принадлежащие к официально-деловому стилю и обладающие так называемым эталонным переводом, выполненным специалистом. Выбор исследуемых сервисов машинного перевода опирался на актуальную статистику запросов пользователей в поисковой системе «Яндекс» [46]. Был проанализирован контекст поисковых запросов, включающих слова «переводчик» и «онлайн-перевод», а также обозначения направления перевода «французский», «с французского на русский», «с русского на французский». Учет запросов с упоминанием какого-либо сервиса машинного перевода обозначил лидерство продуктов компаний Google (более 1,7 млн запросов за месяц), Яндекс (более 1,5 млн запросов за месяц) и PROMT (более 50 тысяч запросов за месяц)[[1]](#footnote-1). Популярность выбранных онлайн-сервисов отражает актуальность и практическую значимость работы. С помощью данных систем МП оригиналы текстов были переведены и вручную выровнены для последующих этапов исследования. Каждому предложению перевода соответствовало эталонное предложение на том же языке. Суммарно было обработано 1134 параллельных контекста. Значительные ограничения на объем исходных текстов накладывала необходимость многократной ручной обработки результатов. Среди дальнейших перспектив развития исследования предполагается увеличение размеров оригинального корпуса.

Этап выделения расхождений в переводах демонстрирует одно из возможных практических решений предобработки текста в анализе ошибок МП (см. Приложение 1). Алгоритм реализован на языке программирования Python с использованием библиотек NLTK [49] и DiffLib [47]. Пакет NLTK (Natural Language Toolkit) представляет собой собрание программ для обработки естественного языка. В нем представлены инструменты, выполняющие токенизацию, лемматизацию, стемминг, частеречную разметку и многие другие этапы анализа текста. Системный модуль DiffLib в скрипте использовался для сравнения последовательностей: получая на входе списки выравненных токенов ручного перевода и перевода, сделанного компьютером, алгоритм выделял найденные несоответствия посредством обозначения их символами «-» и «+». Добавление в алгоритм простого счетчика с условиями «for» и «if», перебирающего каждый элемент списка, начинающийся с символов расхождения DiffLib, позволило рассчитать примерный процент пословных различий машинного и ручного перевода для трех рассматриваемых систем с французского языка на русский язык (см. Рисунок 11) и с русского на французский (см. Рисунок 12).

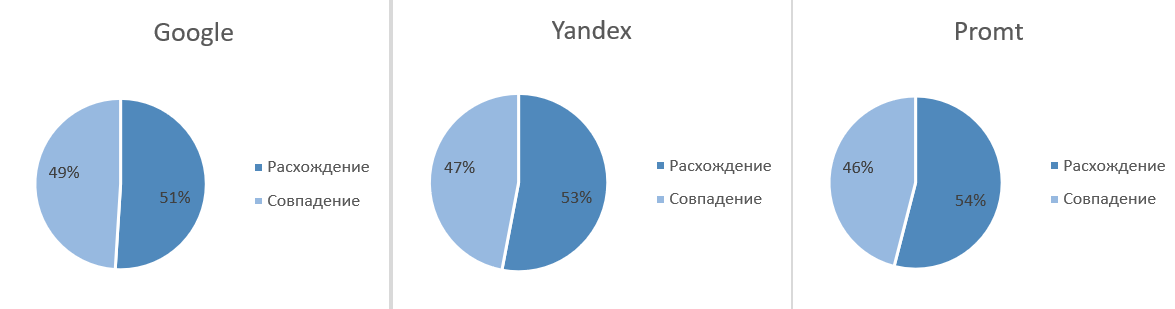


Рисунок 11. Соотношение выделенных автоматически расхождений для результатов работы систем Google, Яндекс и PROMT (FR -> RU)

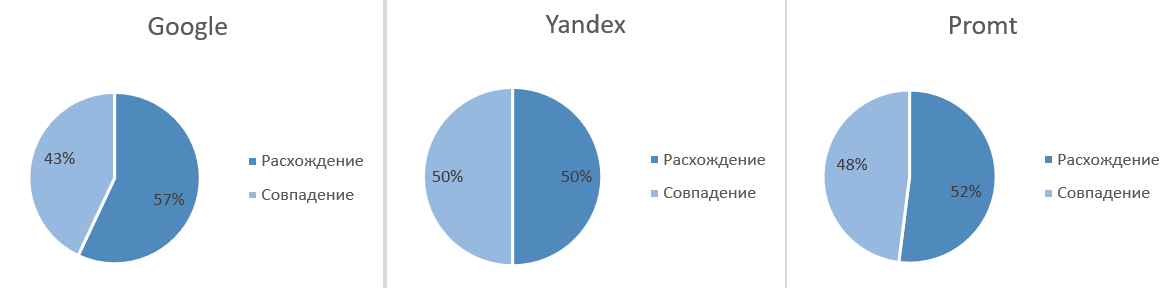


Рисунок 12. Соотношение выделенных автоматически расхождений для результатов работы систем Google, Яндекс и PROMT (RU -> FR)

После этапа автоматической обработки текстов наблюдалось преимущество в количестве совпадений с эталонным переводом у результатов «Google Translate» в направлении «с французского языка на русский» и у результатов «Яндекс - Переводчик» в направлении «с русского языка на французский». Несмотря на то, что построчного сравнения параллельных контекстов было недостаточно для построения классификации, примененное программное решение значительно облегчило этап последующей ручной обработки ошибок.

В самостоятельной обработке несовпадений переводов и составлении классификации был применен следующий подход: каждое выделенное компьютером несоответствие исследовалось с точки зрения релевантности, опускались обнаруженные программой расхождения вследствие грамматически обоснованных перестановок и лексические замены, не несущие искажения смысла фразы. На основании анализа повторно отмеченных несовпадений стало возможно выделить основные типы допускаемых онлайн-сервисами ошибок и детализировать их. В Таблице 2 приведены получившиеся разделы и подразделы составленной таксономии. Каждый вид ошибки снабжался тегом в формате snake\_case с верхним регистром для быстроты осуществления дальнейших подсчетов. Паттерн образования тега представлял собой обозначение раздела типологии ошибок и сокращение названия подраздела.

|  |  |  |
| --- | --- | --- |
| Типы ошибок | | Тег ошибки |
| Лексико-семантические | Неверный перевод аббревиатур | LS\_ABB |
| Неверный перевод имен собственных | LS\_PR\_NAME |
| Нечувствительность к профессиональной лексике и терминологии | LS\_TERM |
| Дословный перевод | LS\_DIR |
| Отсутствие адаптации | LS\_NO\_MOD |
| Искажение смысла слова, неверный подбор синонима | LS\_INACC |
| Синтаксические | Ошибки в построении предложения | SYNT\_STRUCT |
| Ошибки при выражении отношения между словами и частями предложения | SYNT\_REL |
| Ошибки в постановке пунктуационных знаков | SYNT\_PUNCT |
| Морфологические | Ошибки в употреблении предлогов | MORPH\_PREP |
| Ошибки в использовании словоформы | MORPH\_DECL |
| Употребление императива | MORPH\_IMP |
| Стилистические | Использование устаревших норм орфографии | STYL\_OUT |
| Отсутствие единства в обозначениях | STYL\_UN |
| Повторы | STYL\_REP |
| Технические | Частичный перевод | TECH\_PART |
| Неверное распознавание символов | TECH\_TYP |

Таблица 2. Классификация ошибок с подразделением по типам и сопутствующими тегами

После составления классификации производился подсчет обнаруженных ошибок. Суммарно было выделено более 800 различных несовпадений, из них более 300 присутствовало в переводах, выполненных сервисом «PROMT. One». Схожее количество ошибок показали «Google Translate» и «Яндекс – Переводчик»: приблизительно 250 ошибок. Погрешности в подсчетах вызваны частым явлением смешения нескольких типов ошибок в одном контексте. На Рисунке 13 отражено примерное процентное соотношение видов ошибок в зависимости от направления перевода и сервиса, его выполнявшего. При анализе получившегося соотношения прослеживаются некоторые закономерности, поэтому следует подробнее рассмотреть каждый вид неточностей.

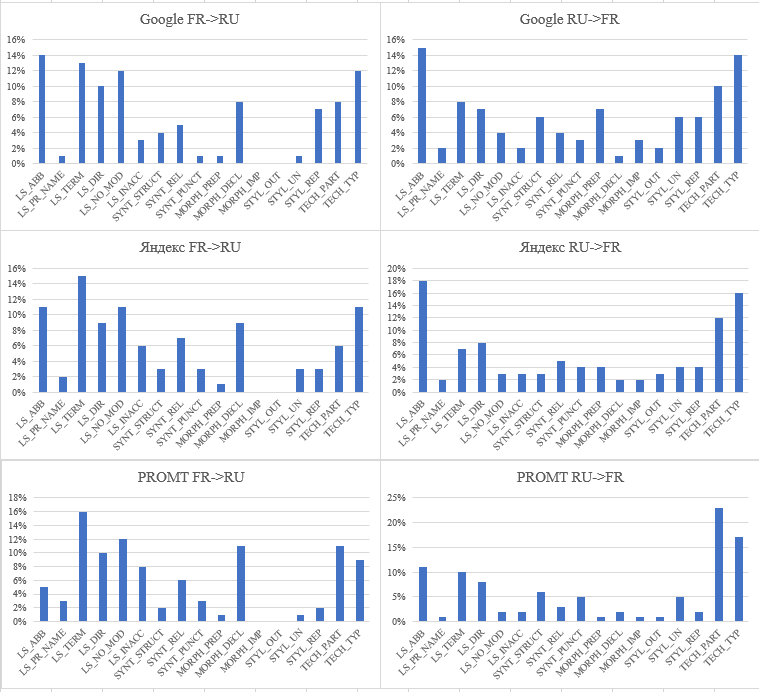


Рисунок 13. Приблизительное процентное соотношение типов ошибок в зависимости от направления перевода и системы МП

* Неверный перевод аббревиатур

При отборе текстов для перевода предполагалось, что частое явление тестов юридической и технической направленности – аббревиатуры – станут «уязвимым местом» систем машинного перевода. Результаты каждого из сервисов это доказывают, однако особое внимание следует обратить на результаты онлайн-переводчика «PROMT. One»: имея меньший процент ошибок в переводе аббревиатур, он «лидирует» в категории частичного перевода, так как чаще всего аббревиатуры либо упускались, либо не переводились данной системой. В обоих направлениях перевода встречаются ошибки в переводах аббревиатур, что влечет за собой неверное согласование глаголов, прилагательных и артиклей. В некоторых случаях аббревиатуры не переводятся вообще, передаются транслитерацией или же переводятся на язык, который не является целевым (английский).

*Примеры:*

**RU->FR:** «АО» («акционерное общество») -> «SPA» («société par actions»);

- Google Translate: «JSC» (вероятно, перевод выполнен только на английский язык – образовано от «joint-stock company»);

- Яндекс – Переводчик: «JSC» (по той же причине);

- PROMT. One: «-», «АО»;

**RU->FR**: «АСК» («автоматизированная система контроля») -> «SCA» («système de contrôle automatisé»);

- Google Translate: «ASK»;

- Яндекс – Переводчик: «ASK»;

- PROMT. One: «АСК»;

**FR->RU:** «SIAE» («système d’information automatisé d’état») -> «ГАИС» («государственная автоматизированная информационная система»);

- Google Translate: «СИАП», «КИС»;

- Яндекс – Переводчик: «АИЭ»;

- PROMT. One: «SIAE»;

* Неверный перевод имен собственных

Среди найденных примеров данного типа ошибок встречается ошибочная транслитерация имен собственных, отсутствие перевода в целом, опущение заглавных букв при переводе, перестановка слов, не требующаяся замена оригинального обозначения на синоним.

*Примеры:*

**RU -> FR:** «Levitin**e** Igor **Evguénievitch**» (ФИО: «Левитин Игорь Евгеньевич»);

- Google Translate: «Igor Levitin»;

- Яндекс – Переводчик: «Igor Levitin Evguenievitch»;

- PROMT. One: «Levitin Igor Evgeniewicz»;

**FR-> RU:** «Président de la Russie V.V. Poutine» -> «Президент России В. В. Путин»;

- Google Translate: **«**Президент России В. В. Путин**»;**

- Яндекс – Переводчик: «**п**резидент России В. В. Путин»;

- PROMT. One: «Президент России **Владимир** Путин **V.V.**»;

* Нечувствительность к профессиональной лексике и терминологии

Зачастую при адаптации профессионализмов требуется более тщательный подбор переводческого эквивалента.

*Примеры:*

**RU -> FR:** «… une entreprise publique russe créée par ordre du Président» («российская государственная компания, созданная по распоряжению Президента»);

- Google Translate: «… une **société d'État** russe créée par ordre du president» (устойчивое сочетание для термина «государственная компания» - **entreprise publique/entreprise d'État**);

- Яндекс – Переводчик: «… une **société d'état** russe créée par ordre du Président» (+ опущение верхнего регистра);

- PROMT. One: «… une **société d « État** russe créée par ordre du président» (+ техническая ошибка - появление символа «, отсутствующего в оригинале, вместо апострофа);

**FR -> RU:** «… создать надежную независимую технологическую базу картографии и геоинформации» («… créer une base technologique indépendante fiable pour la cartographie et la géoinformation»);

- Google Translate: «… создать надежную независимую технологическую базу для **картографирования** и геоинформации»;

- Яндекс – Переводчик: «… создать надежную независимую технологическую базу для **картирования** и геоинформации»;

- PROMT. One: «… создать надежную независимую технологическую базу для картографии и геоинформации»;

Данный пример не иллюстрирует основательное смысловое нарушение дефиниции, но он в полной мере демонстрирует разнообразие синонимического ряда терминологии и сложности выбора подходящего наименования. И. В. Минаева пишет о разнице данных понятий, уточняя, что «картирование» — это общий междисциплинарный термин, обозначающий отображение или моделирование, в то время как «картографирование» отсылает к процессу записи и документирования чего-либо, изображения предметов на карте.

* Дословный перевод

Попытка перевода слова без опоры на семантические связи и контекст чаще всего приводит к ошибке дословного перевода.

*Примеры:*

**RU -> FR:** «… câble 114, bus de mise à la terre 137, 145» («… кабель 114, шины заземления 137, 145»);

- Google Translate: «… câble 114, **jeux de barres** de mise à la terre 137, 145»;

- Яндекс – Переводчик: «… câble 114, **les rails** de mise à la terre 137, 145»;

- PROMT. One: «… **pneumatiques** de terre 137, 145» (если синонимы, приведенные GT и ЯП, еще можно перевести как элемент, соединяющийся с контуром заземления в щитке, то «**pneumatiques**» подразумевает «шины» в значение «резиновые накладки на колеса автомобиля»);

**FR-> RU:** «Основа для разработки многофункциональной интеллектуальной платформы» («La base du développement d'une plateforme intelligente multifonctionnelle»);

- Google Translate: «Основа для разработки многофункциональной интеллектуальной платформы»**;**

- Яндекс – Переводчик: «Основа для разработки многофункциональной интеллектуальной платформы»;

- PROMT. One: «Основа для **развития** многофункциональной интеллектуальной платформы» (обозначение «разработка» в отношении программного обеспечения является более подходящим синонимом, в данном случае перевод был сделан без опоры на контекст);

* Отсутствие адаптации

К этому пункту классификации были отнесены ошибки, заключающиеся в отсутствии использования средств языка перевода для правильной передачи смысла оригинальных фраз.

*Примеры:*

**RU -> FR:** « С’est un environnement unique sécurisé et fiable pour la collecte, le traitement et la transmission des informations» («… является единой защищенной и доверенной средой сбора, обработки и передачи информации»);

- Google Translate: « Le système ERA-GLONASS assure le transfert rapide <…>, **est** un environnement unique sécurisé et fiable pour la collecte, le traitement et la transmission d'informations» (оригинальная конструкция нескольких глагольных сказуемых, разделенных запятой, в французском языке выглядит громоздко и не слишком уместно, логичнее “превратить” часть перечисления в отдельное придаточное предложение («c’est un …»));

- Яндекс – Переводчик: « Le système ERA-GLONASS assure le transfert rapide <…>, **est** un environnement unique sécurisé et fiable pour la collecte, le traitement et la transmission d'informations»;

- PROMT. One: «Le système ERA-GLONASS fournit une communication rapide <…> **et est** un environnement unique et sécurisé pour la collecte, le traitement et la communication des informations»;

* Искажение смысла слова

Периодически в результатах машинного перевода встречаются серьезные искажения смысла переводимого контента.

*Примеры:*

**RU -> FR:** «… adaptateurs 8E273, 8E217, obturateurs 120 (2 pcs.), tuyaux N° 1, N° 2, N° 3» («адаптеры 8E273, 8E217, заглушки 120 (2 шт.), шланги № 1, № 2, № 3»);

- Google Translate: «… adaptateurs 8E273, 8E217, **bouchons** 120 (2 pièces), tuyaux n ° 1, n ° 2, n ° 3»;

- Яндекс – Переводчик: «… adaptateurs 8E273, 8**e**217, **bouchons** 120 (2 pièces), tuyaux n ° 1, n ° 2, n ° 3»;

- PROMT. One: «… adaptateurs 8E273, 8E217, **Les étendues** 120 (2) (слово употребляется в значении «просторы», «широты», возможна какая-то косвенная связь с словом «глушь»), tuyaux n ° 1, n ° 2, n ° 3»;

**FR-> RU:** «За время эксплуатации ГАИС <…>. За все время работы ГАИС <…>» («Pendant l’utilisation du SIAE <…>. Pendant toute la période de fonctionnement du SIAE <…>»);

- Google Translate: «При использовании КИС <…>. За все время работы СИАП <…>»**;**

- Яндекс – Переводчик: «За время эксплуатации АИЭ <…>. За весь период работы АИЭ <…>»;

- PROMT. One: «За время использования **проекта** SIAE <…>. За весь период работы **проекта** SIAE <…>» (+ нарушающее смысл дополнение «проект», не фигурирующее в тексте);

* Ошибка в построении предложений

Основные различия между переводами заключаются в порядке слов в предложениях: в переводе, выполненном специалистом, используется общепринятый порядок слов – добавляющий благозвучности и большей выразительности фразе и расставляющий акценты на важных моментах, в машинном переводе – чаще всего, если это не противоречит правилам целевого языка, сохраняется порядок слов оригинального текста.

*Примеры:*

**FR-> RU:** «В этой связи заинтересованными сторонами **могут выступать** сразу несколько государственных ведомств, ответственных за безопасность на дорогах (аналоги Минтранса, МВД, МЧС, Минздрава) и цифровизацию транспортной отрасли» («À cet égard, plusieurs départements gouvernementaux chargés de la sécurité routière (analogues du ministère des transports, du ministère de l'intérieur, du ministère des situations urgentes, du ministère de la santé) et de la numérisation du secteur des transports peuvent agir en tant que parties prenantes»);

- Google Translate: «В этом отношении несколько государственных ведомств, отвечающих за безопасность дорожного движения (аналогично Министерству транспорта, Министерству внутренних дел, Министерству чрезвычайных ситуаций, Министерству здравоохранения) и оцифровке транспортного сектора, **могут выступать** в качестве заинтересованных сторон»**;**

- Яндекс – Переводчик: «В этой связи в качестве заинтересованных сторон **могут выступать** различные правительственные ведомства, отвечающие за безопасность дорожного движения (аналоги Министерства транспорта, МВД, МЧС, Минздрава) и цифровизации транспортного сектора»;

- PROMT. One: «В связи с этим в качестве заинтересованных сторон **могут действовать** несколько правительственных ведомств, занимающихся вопросами безопасности дорожного движения (аналоги министерства транспорта, МВД, МЧС, Минздрава) и цифровизации транспортного сектора» (может рассматриваться в качестве хорошего примера адаптации синтаксической структуры предложения);

* Ошибки при выражении отношения между словами и частями предложения

Иногда в текстах официально-делового стиля встречаются нагруженные зависимостями предложения, которые тяжело «поддаются» машинному переводу. В данном исследовании был обнаружен достаточно стабильный процент ошибок, связанных с нарушением отношений между словами – примерно 5% в текстах переводов обоих направлений.

*Примеры:*

**RU -> FR:** «Pour construire les systèmes d'information et **de navigation** **pour le** **sauvetage**» («»);

- Google Translate: «Pour construire des systèmes d'information et **de sauvetage de navigation**» (если дословно перевести на русский язык: для создания систем информации и спасения навигации, утеряна смысловая связь);

- Яндекс – Переводчик: «Pour la construction de systèmes d'information et de navigation» (упущен фрагмент);

- PROMT. One: «Pour construire des systèmes d'information et **de navigation de sauvetage**» (практически сохранен смысл);

**FR-> RU:** «Заглушки 120 подключить к устройству 1С227 из состава АСК согласно маркировке соединителей, **предварительно сняв крышки с соединителей**» («raccorder les obturateurs 120 au dispositif 1С227 de la composition du SCA selon le marquage des connecteurs après avoir enlevé les couvercles des connecteurs», в каждом из МП упущен момент предварительного действия);

- Google Translate: «подключить шторки 120 к устройству 1С227 из состава СКА по маркировке разъемов **после снятия крышек с разъемов**»**;**

- Яндекс – Переводчик: «Подсоединить затворы 120 к устройству 1С227 состава SCA по маркировке разъемов **после снятия крышек с разъемов**»;

- PROMT. One: «Соединять затворы 120 в устройстве 1С227 состава SCA согласно маркировке соединителей **подняв крышки соединителей**» (упущен порядок действий, синтаксические зависимости, пунктуация языка перевода);

* Ошибки в постановке пунктуационных знаков

В исследовании были задействованы примеры, в которых наблюдалось отсутствие или избыточность знаков препинания. Чаще всего это было связано с нагруженностью оригинальных синтаксических конструкций.

*Примеры:*

**FR -> RU:** «Деятельность АО «ГЛОНАСС» направлена на создание инновационной информационно-навигационной среды для транспортного комплекса и других отраслей экономики, на предотвращение ущерба жизни и здоровью граждан и окружающей среде» («Les activités de la société anonyme « GLONASS » visent à créer un environnement innovant d'information et de navigation pour le système de transport et pour les autres secteurs de l'économie, afin de prévenir les dommages à la vie et à la santé des citoyens et à l'environnement»);

- Google Translate: «Деятельность ООО «ГЛОНАСС» направлена на создание инновационной информационной и навигационной среды для транспортной системы и других секторов экономики с целью предотвращения нанесения ущерба жизни и здоровью граждан и окружающей среде»;

- Яндекс – Переводчик: «Деятельность ОАО » ГЛОНАСС " направлена на создание инновационной информационно-навигационной среды для транспортной системы и других отраслей экономики**, (= лишняя запятая)** с целью предотвращения ущерба жизни и здоровью граждан и окружающей среде»;

- PROMT. One: «Деятельность анонимной компании "ГЛОНАСС" направлена на создание инновационной информационно-навигационной среды для транспортной системы и для других отраслей экономики с целью предотвращения вреда жизни и здоровью граждан и окружающей среде»;

* Ошибки в употреблении предлогов

Данный раздел включает больше фрагментов перевода с русского на французский. Чаще ошибки связаны с контекстом, в котором используется предлог.

*Примеры:*

**RU -> FR:** «En Fédération de Russie» («В Российской Федерации»);

- Google Translate: «**Dans** la Fédération de Russie» (для стран используется предлог «en»);

- Яндекс – Переводчик: «**Dans** la Fédération de Russie» (та же ошибка);

- PROMT. One: «En Fédération de Russie»;

* Ошибки в использовании словоформы

Свойственно направлению перевода с французского на русский. При изменении композиции предложения теряются зависимости между словами, что в результате приводит к использованию неправильных падежей/чисел/времен и др.

*Примеры:*

**FR-> RU:** «Несколько государственных ведомств, ответственных за безопасность <…> и цифровиза**цию** транспортной отрасли» («Plusieurs départements gouvernementaux chargés de la sécurité <…> et de la numérisation du secteur des transports»);

- Google Translate: «Несколько государственных ведомств, отвечающих за безопасность <…> и оцифро**вке** транспортного сектора»**;**

- Яндекс – Переводчик: «Различные правительственные ведомства, отвечающие за безопасность <…> и цифровиза**ции** транспортного сектора»;

- PROMT. One: «Несколько правительственных ведомств, занимающихся вопросами безопасности <…> и цифровизации транспортного сектора»;

* Употребление императива

Данная ошибка наблюдалась только в направлении RU -> FR: некоторые безличные предложения автоматически были переведены как повелительное наклонение.

*Примеры:*

**RU -> FR:** «Permet de créer une base technologique indépendante fiable» («Позволяет создать надежную независимую технологическую базу»);

- Google Translate: «**Vous permet** de créer une base technologique indépendante fiable pour la cartographie et la geoinformation» («Вам позволено...»);

- Яндекс – Переводчик: «Permet de créer une base technologique indépendante fiable de la cartographie et de l'information géographique»;

- PROMT. One: «Permet de créer une base technologique fiable et indépendante»;

* Использование устаревших норм орфографии

Достаточно интересную закономерность продемонстрировала каждая из платформ: в направлении RU->FR использовалась устаревшая норма французской орфографии [52]. Слово «plateforme» («платформа») с 1990 г. предполагает слитное написание, однако каждый из сервисов использовала вариант «plate-forme». Вероятно, это объясняется большим массивом обучающих данных системы, в котором присутствует именно эта вариация.

* Отсутствие единства в обозначениях

Помимо вышеупомянутых примеров различного перевода и написания аббревиатур, в данную категорию были определены фрагменты перевода с нарушением единообразия в номинациях имен собственных. К примеру, Google Translate в рамках одного предложения переводит название организации «ООН» и как «Nations Unies», и как «ONU».

* Повторы

Такая стилистическая неточность, как повторение, редко выделяется в составе классификаций ошибок МП. Неумение выбрать подходящий по смыслу эквивалент из синонимического ряда является распространенной ошибкой начинающих переводчиков. Однако иногда именно повторы выдают «искусственную природу» текста: неоднократно употребленное слово, не замененное синонимом, бросается в глаза и часто влияет на определение синтаксических зависимостей в фразе.

*Примеры:*

**RU -> FR:** « assure la collecte et le traitement des données de surveillance des véhicules de passagers et de fret transportant des marchandises dangereuses sur tout le territoire de la Russie» («… обеспечивает сбор и обработку данных мониторинга с пассажирского и грузового транспорта, перевозящего опасные грузы на всей территории России»);

- Google Translate: «… assure la collecte et le traitement des données de surveillance du **transport** de passagers et de **marchandises** **transportant** des **marchandises** dangereuses dans toute la Russie»;

- Яндекс – Переводчик: «…assure la collecte et le traitement des données de surveillance des **transports** de passagers et de **marchandises** **transportant** des **marchandises** dangereuses sur tout le territoire de la Russie»;

- PROMT. One: «… assure la collecte et le traitement des données de surveillance à partir du **transport** de passagers et de **marchandises transportant** des **marchandises** dangereuses sur l'ensemble de la Russie»;

**FR-> RU:** «Министерства и их подведомственные организации, а также центральные и местные органы власти» («Ministères et leurs organismes subordonnés, ainsi que les autorités centrales et locales»);

- Google Translate: «Министерства и **подведомственные** им **ведомства**, а также **центральные** и **центральные** органы **власти** и местные **власти**»**;**

- Яндекс – Переводчик: «Министерства и *их подчиненные* ведомства, а также центральные и местные органы власти»;

- PROMT. One: «Министерства и *их подчиненные* ведомства, а также центральные и местные органы власти»;

Отдельной категорией были выделены так называемые «технические» ошибки. В обширный подраздел «Частичный перевод» попали все случаи опущения непереведенного фрагмента, сохранения исходного языка, транслитерации и перевода не на целевой язык. К «Неверному распознаванию символов» были отнесены примеры добавления избыточных знаков, отсутствующих в оригинальном фрагменте и нарушения употребления кавычек.

Таким образом, систематизация видов ошибок и рассмотрение конкретных примеров расхождений позволяют выделить сходства и различия рассматриваемых систем МП. В количественном соотношении ошибок наблюдаются лучшие показатели сервисов компаний Google и Яндекс, однако в некоторых примерах отчетливо прослеживается качество перевода системы PROMT с точки зрения подбора лексики и сохранения содержания. Иногда серьезность смыслового искажения может более негативно сказаться на оценке качества перевода, нежели общее число ошибок.

## ***3.3 Вычисление автоматической оценки качества результатов перевода***

В данном исследовании автоматическая оценка качества перевода производится с целью сопоставления результатов классификации ошибок каждой из систем и вычислений, полученных при помощи статистических метрик. Так проверяется гипотеза о практическом применении автоматической оценки качества перевода в постредактировании: отразят ли формулы результат, полученный в ходе ручного анализа эталонных и целевых текстов.

В качестве программного решения были использованы библиотека NLTK, в которой реализованы BLEU, NIST и METEOR, и пакет PyRouge, содержащий метрику ROUGE. Метрики BLEU и NIST в качестве hypothesis и reference принимали токенизированные тексты, поэтому токенизация также производилась при помощи модуля word\_tokenize в NLTK, аналогично этапу предобработки текстов. Результаты вычислений представлены в Таблице 3.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Система / направление | BLEU | NIST | METEOR | ROUGE |
| Google (RU -> FR) | 0.6415245 | 7.7078529 | 0.7540957 | 0.5981532 |
| Яндекс (RU -> FR) | 0.5209114 | 6.9313279 | 0.6605091 | 0.5700889 |
| PROMT (RU -> FR) | 0.4148033 | 6.4116869 | 0.5676166 | 0.5423234 |
| Google (FR -> RU) | 0.4506869 | 6.5719088 | 0.5791645 | 0.4785456 |
| Yandex (FR -> RU) | 0.4183229 | 6.3083456 | 0.5674998 | 0.3445123 |
| PROMT (FR -> RU) | 0.4389211 | 6.5562004 | 0.6154441 | 0.3834456 |

Таблица 3. Результаты автоматической оценки качества результатов работы сервисов онлайн-перевода компаний Google, Яндекс и PROMT

В качестве интерпретации полученных значений можно сделать следующие выводы:

* Прослеживается динамика повышения значений f-меры каждой из метрик в направлении перевода с русского языка на французский;
* Алгоритм BLEU демонстрирует лидерство сервиса «Google Translate» в обоих направлениях перевода: с русского на французский язык – абсолютное лидерство и достаточно высокое значение f-меры, с французского на русский язык – относительное лидерство (разница с сервисом от PROMT составляет менее 0,02) и среднее значение f-меры;
* Алгоритм NIST также выделяет Google Translate в направлении RU -> FR. В обратном направлении показатели метрики NIST Google Translate и PROMT. One примерно равны;
* По результатам вычисления метрики METEOR онлайн-переводчик компании PROMT лидирует в направлении «с французского на русский»;
* Результаты метрики ROUGE, аналогично алгоритму BLEU, подчеркивают преимущество системы Google Translate;

## ***3.4 Вычисление автоматической оценки читабельности результатов перевода***

Завершающим этапом практической части исследования стало вычисление читабельности результатов перевода. Для достижения более точных результатов в качестве материала использовались тексты объемом 40 предложений. В программном решении был использован вышеупомянутый механизм токенизации текстов. Для подсчета слогов в метриках SMOG и F-K была написана отдельная функция, принимающая значения гласных русского и французского алфавитов. Скрипт вычисления формул удобочитаемости был написан самостоятельно с использованием модуля Math, включающего разнообразные математические операции. Полученные значения для удобства сравнения были округлены до двух знаков после запятой (см. Таблица 4).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Система / направление | ARI | F-K | SMOG | C-L |
| Эталон (FR) | 17.96 | 19.86 | 20.53 | 15.58 |
| Google (RU -> FR) | 18.65 | 19.77 | 20.32 | 17.16 |
| Яндекс (RU -> FR) | 19.51 | 20.12 | 20.56 | 17.43 |
| PROMT (RU -> FR) | 19.07 | 19.89 | 20.16 | 17.85 |
| Эталон (RU) | 21.98 | 24.7 | 21.19 | 24.66 |
| Google (FR -> RU) | 21.91 | 24.44 | 20.88 | 25.3 |
| Yandex (FR -> RU) | 22.92 | 25.22 | 22.03 | 25.23 |
| PROMT (FR -> RU) | 22.83 | 24.96 | 21.65 | 25.48 |

Таблица 4. Результаты автоматической оценки читабельности результатов работы сервисов онлайн-перевода компаний Google, Яндекс и PROMT

В ходе анализа полученных значений индексов можно сделать вывод о применимости данных формул в оценке качества машинного перевода. Некоторые формулы действительно демонстрируют разницу в результатах, опираясь на которую, можно сделать вывод о читаемости результатов работы исследуемых сервисов: стабильно высокие показатели сложности текста наблюдаются у «Яндекс – Переводчика», «Google Translate» периодически способен выдавать текст, читаемый легче, чем эталонный перевод, а «PROMT. One» по многим метрикам занимает второе по сложности синтезируемых текстов место. Интересным наблюдением является то, что тексты, переведенные с русского на французский, определяются метриками, как значительно более читабельные.

## ***3.5 Выводы по Главе 3***

В результате практического исследования была составлена собственная классификация ошибок машинного перевода языковой пары «русский –французский», посчитаны автоматические оценки качества результатов машинного перевода и читабельности текстов, генерируемых онлайн-сервисами «Google Translate», «Яндекс – Переводчик» и «PROMT. One». В составлении классификации были применены автоматическое выделение расхождений в переводах и метод разметки выделенных несоответствий специальными тегами для унификации обозначений.

Для оценки удобочитаемости и качества переводов были применены популярные метрики, описанные в Главе 2. В качестве программных решений использовались скрипты на языке Python с использованием библиотек, предназначенных для обработки текстов на естественном языке.

# ***Заключение***

Данное исследование содержит результаты сопоставительного анализа результатов машинного перевода, выполненного онлайн-сервисами «Google Translate», «Яндекс – Переводчик» и «PROMT. One» для языковой пары «русский - французский». В ходе работы были рассмотрены понятия, связанные с автоматическим переводом, история развития научного направления и особенности существующих систем машинного перевода. Отдельное внимание уделялось направлениям перевода «с русского языка на французский» и «с французского языка на русский» в контексте переводоведения. Методы оценки результатов машинного перевода были проанализированы с точки зрения возможности их применения в практике переводоведения. В рамках работы был произведен сравнительный анализ результатов перевода русских и французских текстов, разработана форма классификации ошибок МП, разобраны примеры ошибок каждой из рассматриваемых систем, вычислены автоматические оценки качества и читабельности переводов. В результате всех этапов исследования можно сделать следующие выводы:

1. Использованный в работе метод автоматической предобработки текстов для выделения несоответствий в переводах может быть применен в составлении классификаций ошибок других направлений перевода и в поиске примеров неточностей МП для обучающих пособий и курсов по автоматизированному переводу.
2. Использование тегов для выделенных несоответствий в переводах значительно облегчает процесс ручной разметки.
3. Найденные примеры ошибок МП могут быть использованы в качестве тестов для проверки знаний постредактора.
4. Анализ обнаруженных ошибок в результатах работы систем машинного перевода продемонстрировал преимущество сервисов «Google Translate» и «Яндекс – Переводчик». Однако в некоторых приведенных примерах прослеживается высокое качество работы сервиса «PROMT. One» с точки зрения адаптации перевода, лексического разнообразия и смыслового соответствия оригиналу.
5. Автоматическая оценка качества перевода, несмотря на сложность интерпретации полученных результатов, демонстрирует результаты, аналогичные выводам, сделанным в процессе составления классификации. Самую высокую оценку получила система компании Google. Вычисления, полученные при помощи разных метрик, обладали схожими значениями, что позволило сопоставить их и сделать общие выводы.
6. Подсчет читабельности результатов МП с использованием метрик ARI, FK, SMOG, CL может найти практическое применение в оценке трудозатрат постредактора.
7. В качестве возможных перспектив исследования представляется релевантным расширение корпуса текстов, дополнение составленной классификации, анализ причин возникновения ошибок машинного перевода и поиск вариантов их устранения.

# ***Список использованных источников***

***Литература:***

1. *Александрова Е. М.* Влияние лексического сходства языков на переводимость каламбура // Филологические науки. Вопросы теории и практики. 2019. №4. С. 108-113.
2. *Андреева, А. Д.* Обзор систем машинного перевода // Молодой ученый. 2013. № 12. С. 64-66.
3. *Апресян Ю. Д., Богуславский И. М., Иомдин Л. Л., Лазурский А. В., Митюшин Л. Г., Санников В. З., Цинман Л. Л.* Лингвистический процессор для сложных информационных систем. М.: Наука, 1992.
4. *Бархударов Л. С.* Язык и перевод (Вопросы общей и частной теории перевода). М.: «Междунар. отношения», 1975.
5. *Белобородова Е.А., Бодулева А.Р.* Преимущества и недостатки машинного перевода // Инновационная наука. 2016. №9 (21). С. 224-226.
6. *Бунтман Н. В., Гончаров А. А., Нуриев В. А.* Ошибки в машинном переводе: проблемы классификации // Системы и средства информатики. 2019. Т. 29. № 3. С. 92-103.
7. *Владова И. М.* Перевод как межкультурная коммуникация // Русский язык за рубежом. 2011. № 4(227). С. 27-32.
8. *Гак В. Г*. Сопоставительная лексикология: на материале французского и русского языков. М.: «Междунар. отношения», 1977.
9. *Гриф М. Г., Тимофеева М. К.* Интерлингва в системах машинного перевода для жестовых языков // Труды СПИИРАН. 2012. № 20 (1). C. 116-137.
10. *Добровольский Б. Д.* Лексические трудности перевода в лингвокультурном аспекте. М., 2009.
11. *Дроздова К.А.* Машинный перевод: история, классификация, методы // Вестник Омского государственного педагогического университета. Гуманитарные исследования. 2015. №3 (7). С. 156-158.
12. *Замуруева H. A*. Жанровые особенности официально-делового стиля речи // Ученые записки ОГУ. Серия: Гуманитарные и социальные науки. 2010. №3-2. С. 120-123.
13. *Ивтушок Е.* Машинный перевод не справился с пониманием больших текстов. 2018. URL: <https://nplus1.ru/news/2018/09/05/machine-translation-evaluation> (дата последнего обращения: 01.06.2021).
14. *Карцева Е. Ю., Маргарян Т. Д., Гурова Г. Г.* Развитие машинного перевода и его место в профессиональной межкультурной коммуникации // Вестник РУДН. Серия: Теория языка. Семиотика. Семантика. 2016. №3. С. 155-164.
15. *Кенжаев А. Д.* Машинный перевод: история и современность // Иностранные языки и регионоведение. М.: 2014.
16. *Кольцова, Д. А., Кольцов С. В.* История и развитие машинного перевода // Русский язык и культура в зеркале перевода. 2019. № 1. С. 130-135.
17. *Комиссаров В. Н.* Теория перевода. М., 1990. Т. 8.
18. *Кочеткова Н.* *С., Ревина Е. В.* Особенности машинного перевода // Филологические науки. Вопросы теории и практики. Тамбов: Грамота, 2017. № 6. Ч. 2. С. 106-109.
19. *Кулагина О. С.* О современном состоянии машинного перевода // Математические вопросы кибернетики. Вып. 3. М.: Наука, 1991. С. 5–50.
20. *Кушеева М.Н., Гомбожапова Т.Н., Аюшеева Н.Н.* Метод оценки качества семантической сети текста // Современные наукоемкие технологии. 2019. № 11-1. С. 77-81.
21. *Легостина М. С.* Метрики оценки качества машинного перевода // Инноватика-2019: сборник материалов XV Международной школы-конференции студентов, аспирантов и молодых ученых. Томск: Общество с ограниченной ответственностью "СТТ". 2019. С. 490-493.
22. *Люльчева Е. М.* Имплицитные способы выражения отрицания // Вестник ИГЛУ. 2013. №2 (23). С. 100-108.
23. *Минаева И. В.* Картирование как метод репрезентации в работах Ойвинда Фальстрема // Артикульт. 2014. 16(4). С. 104-113.
24. *Митренина О. В.* Машинный перевод // Прикладная и компьютерная лингвистика. 2-е изд. ред. М.: Издательская группа URSS. 2017. С. 156-189.
25. *Мифтахова Р. Г.* Технологии машинного перевода. Нейроперевод // Доклады Башкирского университета. 2018. Т. 3. № 6. С. 711-715.
26. *Мохов Н. А.* Новейшие информационные технологии в переводе: опыт и перспективы // Язык и культура: Сборник статей XXIX Международной научной конференции. Томск: Национальный исследовательский Томский государственный университет. 2019. С. 230-235.
27. *Рогушина Ю. В*. Использование критериев оценки удобочитаемости текста для поиска информации, соответствующей реальным потребностям пользователя // Проблеми програмування. 2007. № 3. C. 76–87.
28. *Рычихин А. К.* О методах оценки качества машинного перевода // Системы и средства информ. 2019. Т. 29. № 4. С. 106–118.
29. *Солнышкина М. И., Гафиятова Э. В., Казачкова М. Б.* Инструменты оценки сложности текста: современная парадигма // Научное наследие Е.И. Пассова в контексте развития иноязычного образования: Материалы Международной научной конференции, посвященной 90-летию со дня рождения Е.И. Пассова, Липецк, 01–02 декабря 2020 года / Под общей редакцией В.Б. Царьковой, А.А. Люлюшина. Липецк: Липецкий государственный педагогический университет имени П.П. Семенова-Тян-Шанского, 2020. С. 251-263.
30. *Томилова А. И.* Явление межъязыковой псевдоэквивалентности в русском и французском языках: теоретические и прикладные аспекты. Екатеринбург: Уральский государственный педагогический университет (УрГПУ), 2018.
31. *Томин В. В.* О проблемах машинного перевода научно-технического текста в информационном поле кросс-культурного взаимодействия // Вестник ОГУ. 2015. №1 (176). С. 33-39.
32. *Федюк П. С.* Специфика определения критериев широкозначности лексических единиц // Известия РГПУ им. А. И. Герцена. 2009. №101. 200-205.
33. *Худяков Н. А.* Современные технологии машинного перевода текстов: преимущества и недостатки // Иностранные языки: лингвистические и методические аспекты. 2019. № 41. С. 208-214.
34. *Якобсон Р. О.* О лингвистических аспектах перевода // Вопросы теории перевода в зарубежной лингвистике. М., 1978. С. 16-24.
35. *Banerjee S., Lavie A.* METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments // Proceedings of Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization at the 43rd Annual Meeting of the Association of Computational Linguistics (ACL-2005). Michigan, 2005.
36. *Bromley A.* Charles Babbage's Analytical Engine, 1838 // IEEE Annals of the History of Computing. T. 4. № 3. 1982. P. 196-217.
37. *Carl M.* Post-Editing Machine Translation: Efficiency, Strategies, and Revision Processes in Professional Translation Settings // Psycholinguistic and Cognitive Inquiries into Translation and Interpreting. 2015. P. 145-174.
38. *Cheragui M.* Theoretical overview of machine translation // CEUR Workshop Proceedings. № 867. 2012. P. 160-169.
39. *Hutchins J.* Machine Translation: A Brief History. Oxford: Pergamon Press, 1995.
40. *Koponen M.* Machine Translation Post-editing and Effort. Empirical Studies on the Post-editing Process. Helsinki, 2016.
41. *Lin Ch.* ROUGE: A Package for Automatic Evaluation of summaries. Proceedings of the ACL Workshop: Text Summarization Braches Out. 2004.
42. *Papineni K., Roukos S., Ward T., Zhu W.* BLEU: A Method for Automatic Evaluation of Machine Translation. 2002.
43. *Peng H.* The Impact of Machine Translation and Computer-aided Translation on Translators // IOP Conference Series: Materials Science and Engineering. 2018.
44. *Zhou S., Jeong H., Green P.* How consistent are the best-known readability equations in estimating the readability of design standards? // IEEE Transactions on Professional Communication. 2017.
45. *Zong Zh.* Research on the Relations Between Machine Translation and Human Translation. Journal of Physics: Conference Series. 2018.

***Интернет-ресурсы****:*

1. Яндекс – Переводчик. URL: <https://translate.yandex.ru/>. (Дата последнего обращения: 01.06.2021).
2. Яндекс Подбор Слов. URL: <https://wordstat.yandex.ru/>. (Дата последнего обращения: 01.06.2021).
3. DiffLib. URL: <https://docs.python.org/3/library/difflib.html>. (Дата последнего обращения: 01.06.2021).
4. Google Translate. URL: <https://translate.google.com/>. (Дата последнего обращения: 01.06.2021).
5. Natural Language Toolkit. URL: <https://www.nltk.org/>. (Дата последнего обращения: 01.06.2021).
6. PROMT One. URL: <https://www.translate.ru/>. (Дата последнего обращения: 01.06.2021).
7. RapidAPI. Top 8 Best Translation Websites & Services (in 2021). URL: <https://rapidapi.com/blog/best-translation-sites/>. (Дата последнего обращения: 01.06.2021).
8. RENOUVO (Réseau pour la nouvelle orthographe du français). URL: <http://www.renouvo.org/liste.php>. (Дата последнего обращения: 01.06.2021).

# ***Приложение 1***

***Листинг программы для предобработки и выделения расхождений в переводах***

import difflib

import nltk

with open('file\_name1.txt', "r", encoding="utf-8") as file\_1:

file\_1\_text = file\_1.read()

with open('file\_name2.txt', "r", encoding="utf-8") as file\_2:

file\_2\_text = file\_2.read()

file\_1\_sent = nltk.tokenize.sent\_tokenize(file\_1\_text)

file\_2\_sent = nltk.tokenize.sent\_tokenize(file\_2\_text)

file\_1\_word = nltk.tokenize.word\_tokenize(file\_1\_text)

file\_2\_word = nltk.tokenize.word\_tokenize(file\_2\_text)

diff\_sentences = []

diff\_words = []

# Find and print the diff:

for line in difflib.unified\_diff(

file\_1\_sent, file\_2\_sent, fromfile='file1.txt',

tofile='file2.txt', lineterm=''):

diff\_sentences.append(line)

for line in difflib.unified\_diff(

file\_1\_word, file\_2\_word, fromfile='file1.txt',

tofile='file2.txt', lineterm=''):

diff\_words.append(line)

print(diff\_sentences)

print(diff\_words)

# ***Приложение 2***

***Листинг программы для автоматической оценки качества перевода посредством метрик BLEU, NIST, METEOR и ROUGE***

import nltk

import pyrouge\_master

from pyrouge\_master.PyRouge import pyrouge

from pyrouge\_master.PyRouge.pyrouge import Rouge

bleu\_score = nltk.translate.bleu\_score.sentence\_bleu()

nist\_score = nltk.translate.nist\_score.sentence\_nist()

meteor\_score = nltk.translate.meteor\_score.meteor\_score()

rouge\_score = Rouge.rouge\_l()

with open('file\_name1.txt', "r", encoding="utf-8") as file\_1:

file\_1\_text = file\_1.read()

with open('file\_name2.txt', "r", encoding="utf-8") as file\_2:

file\_2\_text = file\_2.read()

file\_1\_word = nltk.tokenize.word\_tokenize(file\_1\_text)

file\_2\_word = nltk.tokenize.word\_tokenize(file\_2\_text)

print(bleu\_score([file\_2\_word], file\_1\_word))

print(nist\_score([file\_2\_word], file\_1\_word))

print(meteor\_score([file\_2\_text], file\_1\_text))

print(rouge\_score([file\_2\_text], [file\_1\_text])[2])

# ***Приложение 3***

***Листинг программы для подсчета метрик читабельности текста***

import nltk

import string

import math

def syllable\_count(word):

word = word.lower()

count = 0

vowels = "аоуэыюяиееeéêèaâioôuy"

if word[0] in vowels:

count += 1

for index in range(1, len(word)):

if word[index] in vowels and word[index - 1] not in vowels:

count += 1

if word[index] in vowels and word[index - 1] in vowels and word[index - 2] in vowels:

count += 1

if count == 0:

count += 1

return count

def round\_score(score):

return round(score, 2)

with open('file\_name1.txt', "r", encoding="utf-8") as file:

text = file.read()

sentences = nltk.tokenize.sent\_tokenize(text)

sentences\_num = len(sentences)

all\_words\_num = 0

all\_letters\_num = 0

all\_syl\_num = 0

all\_pol\_syl\_num = 0

for sent in sentences:

words = nltk.tokenize.word\_tokenize(sent)

words = [i.lower() for i in words if (i not in string.punctuation)]

words\_num = len(words)

for word in words:

syl\_num = syllable\_count(word)

all\_syl\_num += syl\_num

if syl\_num > 2:

all\_pol\_syl\_num += 1

letters = re.findall(r'[\S]', sent)

letters\_num = len(letters)

all\_words\_num += words\_num

all\_letters\_num += letters\_num

api\_score = 4.71 \* all\_letters\_num / all\_words\_num + 0.5 \* all\_words\_num / sentences\_num - 21.43

fk\_score = 0.39 \* all\_words\_num / sentences\_num + 11.8 \* all\_syl\_num / all\_words\_num - 15.59

smog\_score = 1.043 \* math.sqrt(all\_pol\_syl\_num \* 30 / sentences\_num) + 3.1291

l = all\_letters\_num / all\_words\_num \* 100

s = sentences\_num / all\_words\_num \* 100

cl\_score = 0.0588 \* l - 0.296 \* s - 15.8

average = (round\_score(api\_score) + round\_score(fk\_score) + round\_score(smog\_score) + round\_score(cl\_score)) / 4

print('Words: {}\nSentences: {}\nCharacters: {}\nSyllables: {}\nPolysyllables: {}'.format(all\_words\_num, sentences\_num, all\_letters\_num, all\_syl\_num, all\_pol\_syl\_num))

print('Enter the score you want to calculate (ARI, FK, SMOG, CL, all):')

formula = input()

if formula == 'ARI':

print('Automated Readability Index: {}.'.format(round\_score(api\_score)))

if formula == 'FK':

print('Flesch–Kincaid readability tests: {}.'.format(round\_score(fk\_score)))

if formula == 'SMOG':

print('Simple Measure of Gobbledygook: {}.'.format(round\_score(smog\_score)))

if formula == 'CL':

print('Coleman–Liau index: {}.'.format(round\_score(cl\_score)))

if formula == 'all':

print('Automated Readability Index: {}.'.format(round\_score(api\_score)))

print('Flesch–Kincaid readability tests: {}.'.format(round\_score(fk\_score)))

print('Simple Measure of Gobbledygook:{}.'.format(round\_score(smog\_score)))

print('Coleman–Liau index: {}.'.format(round\_score(cl\_score)))

print('Average index: {}.'.format(average))

1. На момент мая 2021 г. [↑](#footnote-ref-1)